Discovering Motion Primitives for Unsupervised Grouping and One-shot Learning of Human Actions, Gestures, and Expressions

Yang Yang, Imran Saleemi, and Mubarak Shah

Abstract—This paper proposes a novel representation of articulated human actions and gestures, and facial expressions. The main goals of the proposed approach are: (1) to enable recognition using very few examples, i.e., one, or k-shot learning, and (2) meaningful organization of unlabelled data sets by unsupervised clustering. Our proposed representation is obtained by automatically discovering high level sub-actions or motion primitives, by hierarchical clustering of observed optical flow in four dimensional, spatial and motion flow space. The completely unsupervised proposed method, in contrast to state of the art representations like bag of video words, provides a meaningful representation conducive to visual interpretation and textual labeling. Each primitive action depicts an atomic sub-action, like directional motion of limb or torso, and is represented by a mixture of four dimensional Gaussian distributions. For one-shot and k-shot learning, the sequence of primitive labels discovered in a test video are labelled using KL divergence, and can then be represented as a string and matched against similar strings of training videos. The same sequence can also be collapsed into a histogram of primitives, or be used to learn a Hidden Markov model to represent classes. We have performed extensive experiments on recognition by one and k-shot learning as well as unsupervised action clustering on six human actions and gesture datasets, a composite dataset, and a database of facial expressions. These experiments confirm the validity, and discriminative nature of the proposed representation.

Index Terms—human actions, one-shot learning, unsupervised clustering, gestures, facial expressions, action representation, action recognition, motion primitives, motion patterns, histogram of motion primitives, motion primitives strings, Hidden Markov model

1 Introduction

Learning using few labeled examples should be an essential feature in any practical action recognition system because collection of a large number of examples for each of many diverse categories is an expensive and laborious task. Although humans are adept at learning new object and action categories, the same cannot be said about most existing computer vision methods, even though such capability is of significant importance. A majority of proposed recognition approaches require large amounts of labeled training data, while testing using either a leave-one-out or a train-test split scenario. In this paper, we put forth a discriminative yet flexible representation of gestures and actions, that lends itself well to the task of learning from few as possible examples. We further extend the idea of one-shot learning to attempt a perceptual grouping of unlabelled datasets and to obtain subsets of videos that correspond to a meaningful grouping of actions, for instance, recovering the original class-based partitions.

Given the very large volume of existing research in the area of action recognition, we observe that action representation can range from spatiotemporally local to global features. On one end of this spectrum are interest-point based representations where a single descriptor encodes the appearance [27] or motion [22], [8] in very small $x-y-t$ volumes, while on the other hand features based on actor silhouettes [7], contours [50], and space-time volumes [36] attempt to capture the entire action in a single feature ignoring the intuitive compositional hierarchy of action primitives. Some of the recent approaches which have performed well on standard datasets [43],

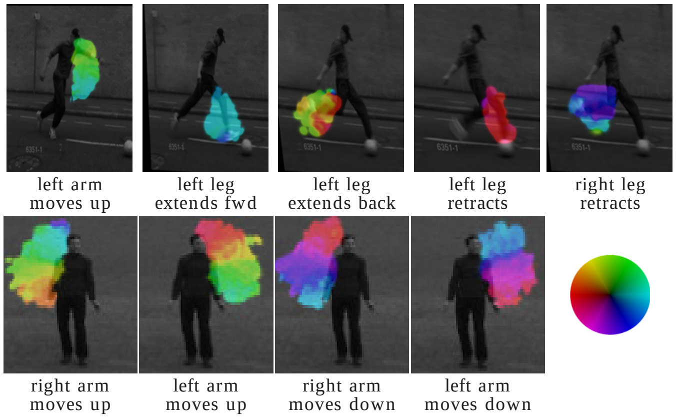


Fig. 1. Proposed probabilistic representation of primitive actions. Top: five primitives discovered in UCF Sports ‘kicking’ action, showing relative location, shape, and flow of the sub-actions. Colors correspond to conditional expected flow magnitude and direction given pixel locations, as per the color wheel. Bottom: 4 primitives for the KTH ‘waving’ action.

- *Y. Yang, I. Saleemi, and M. Shah are with the Computer Vision Lab, Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL, 32816. E-mail: {yyang, imran, shah}@eecs.ucf.edu*

as well as relatively older work involving manual steps [30], [51] tend to lie between these two extremes in terms of the abstract spatiotemporal scale at which a video is represented.

This observation forms the basis of the proposed representation with the underlying idea that intermediate features (action primitives) should: (a) span as large as possible but contiguous $x-y-t$ volumes with smoothly varying motion, and (b) should be flexible enough to allow deformations arising from articulation of body parts. A byproduct of these properties is that the intermediate representation will be conducive to human understanding. In other words, a meaningful action primitive is one which can be illustrated visually, and described textually, e.g., 'left arm moving upwards', or 'right leg moving outwards and upwards', etc. We argue and show experimentally, that such a representation is much more discriminative, and makes the tasks of 'few-shot' action, gesture, or expression recognition, or unsupervised clustering simpler as compared to traditional methods. This paper proposes such a representation based on motion primitives, examples of which are shown in Fig. 1. A summary of our method to obtain the proposed representation follows.

1.1 Algorithmic Overview

An algorithmic overview of the proposed approach is as illustrated in Fig. 2: given a video containing an action, (i) when required, camera motion is compensated to obtain residual actor-only motion, (ii) a frame difference based foreground estimation, and 'centralization' of the actor to remove translational motion is performed, thus resulting in a stack of rectangular image regions coarsely centered around the human; (iii) computation of optical flow to obtain 4d feature vectors (x, y, u, v); (iv) clustering of feature vectors to obtain components of a Gaussian mixture; (v) spatio-temporal linking of Gaussian components resulting in instances of primitive actions; and (vi) merging of primitive action instances to obtain final statistical representation of the primitives.

For supervised recognition, given a test video, instances of action primitives are detected in a similar fashion, which are labeled by comparing against the learned primitives. Sequences of observed primitives in training and test videos are represented as strings and matched using simple alignment [28] to classify the test video. We also experimented with representation of primitive sequences as histograms, followed by classifier learning, as well as using temporal sequences of primitive labels to learn state transition models for each class.

Compared to the state of the art action representations the contributions of the proposed work are:

- Completely unsupervised discovery of representative and discriminative action primitives without assuming any knowledge of the number of primitives present, or their interpretation,
- A novel representation of human action primitives that captures the spatial layout, shape, temporal extent, as well as the motion flow of a primitive,
- Statistical description of primitives as motion patterns, thus providing a generative model, capable of estimating confidence in observing a specific motion at a specific point in space-time, and even sampling,
- Highly abstract, discriminative representation of primitives which can be labeled textually as components of an action, thus making the recognition task straightforward.

In the following section, we present a brief review of the most relevant representation and recognition techniques in the literature. We then describe the details of the proposed approach for primitives discovery in Section 3, and our representation for actions, gestures, and facial expressions in Section 4. Experiments and results are reported in Section 5.

2 Related Work

Human action and activity recognition is a broad and active area of research in computer vision, and comprehensive reviews of the proposed methods can be found in [42], [48]. Our discussion in this regard is restricted to a few influential and relevant parts of literature, with a focus on representation, as compared to machine learning and classification approaches. These methods can be categorized based on the different levels of abstraction at which human actions are represented, and are summarized below:

Interest point based Representations: One important direction of research in the human action literature which has gained a lot of interest recently is the use of spatiotemporal interest points, and feature descriptors or trajectories computed around those interest points. Works by Dollar et al. [8], Laptev et al. [22], Gilbert et al. [12], and Filipovych and Ribeiro [11] are representative of this large category of methods, which is also loosely termed as 'bag of visual or video words'. Many of the state of the art approaches in action recognition, like [43], fall in this category. The main strength of this representation is the robustness to occlusion, since there is no need to track or detect the entire human body or its parts, and therefore, impressive results have been obtained using such methods on standard data sets. The same strengths of this model however, make it less than ideally suited to content understanding and textual descriptions, mainly because it is too local (visual words span very small spatiotemporal regions) and too distributed (histogram ignores spatial and temporal ordering of words). Indeed more recent approaches [43] attempt to mitigate the former of the above two.

Moreover, these methods are not exempt from sensitivity to a number of intermediate processes including interest point detection, choice of descriptors, number of codebook words, and classifiers. A large number of methods have been proposed to address each of these

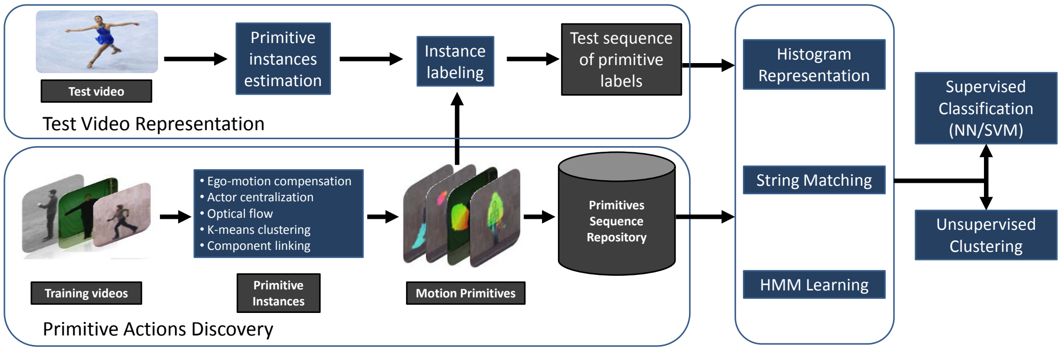


Fig. 2. Process flow of the proposed approach for action representation, recognition, and clustering.

problems. We contend that action representation and recognition need not be this complex and should be visually and textually interpretable, and especially so when the goal is content understanding rather than state of the art quantitative performance.

Holistic Representations: These methods incorporate global, image-based representation of actions, and do not require detection, labeling, or tracking of individual body parts. The only requirement for these methods is to detect a bounding box, or contour enclosing the person performing an action. Dense features computed within the bounding box are then used to represent actions. Examples of these features include intensity images [6], silhouettes composing motion history (MHI) and motion energy (MEI) images [3], spatiotemporal shapes using tracks of body contours [50], and spatiotemporal volumes spanned by silhouette images [36]. Other holistic representations of actions involve optical flow [32], spatiotemporal gradients [52], and HoG [39]. As mentioned earlier, such representations ignore the useful information related to the primitive sub-actions, which can compose multiple actions by spatiotemporal ordering, and are much more flexible than holistic representations. Performance of holistic representations is expected to drop as diversity and noise within examples of a class increases, since these are rigid and brittle.

Part based Representations: Methods based on information derived from knowledge of location, or appearance of body parts or joints fall in this category. This is the most intuitive representation, but the most difficult to estimate reliably in practice. Examples include body parts detection, or features of trajectories of landmark human body joints [51], and motion features of detected body parts [30]. Detection of body parts or joints in videos is an extremely challenging problem and even the constrained settings of discriminative background and use of markers does not ensure a completely unsupervised process.

Other examples of work in this category include Ke et al [16] who proposed the learning of a cascade of filters using space-time volumetric features, effectively performing action detection as well as precise spatiotemporal localization. In [17], over-segmented video volumes without regards to actor parts, are matched to volumetric representation of an event using shape and flow cues in order to perform detection of actions in crowded scenes. Singh and Nevatia [37] put forth a joint tracking and recognition approach, where they learn action models by annotating key poses in 2D, and propose an approach for pose localization using a pose specific part model. Their approach was tested on two gesture datasets. Tran et al [41] have also recently proposed to model relative motion of body parts for action recognition.

Our proposed work can be considered a part of this category, since the proposed primitive action representation *generally* corresponds to discriminative motion induced by independently moving body parts. In contrast to traditional methods however, there is no need to explicitly detect any body parts, or even assume presence of such parts or joints. The primitives correspond to any large spatiotemporal regions of contiguous, continuous, and smooth flow.

In light of this discussion, we now describe the proposed action representation, which is completely unsupervised, discriminative, and simplifies the tasks of action recognition, classification, or clustering.

3 Motion Primitives Discovery

The goal of the proposed human action representation is twofold: (i) to automatically discover discriminative, and meaningful sub-actions (or primitives) within videos of articulated human actions, without assuming priors on their number, type, or scale, and (ii) to learn the parameters of a statistical distribution that best describes the location, shape, and motion flow of these primitives, in addition to their probability of occurrence. The idea for the proposed representation is inspired by several recent works in traffic patterns discovery and representation [45], [20], [34]. However, instead of human actions, these techniques have been proposed for learning high level events and activities in surveillance scenarios, by observing rigid body motion of objects, like vehicles, over long periods of time.

Since our choice for action primitives modeling is to estimate a statistical description of regions of similar optical flow, we draw from the method of [34], which learns Gaussian mixture distributions to estimate traffic

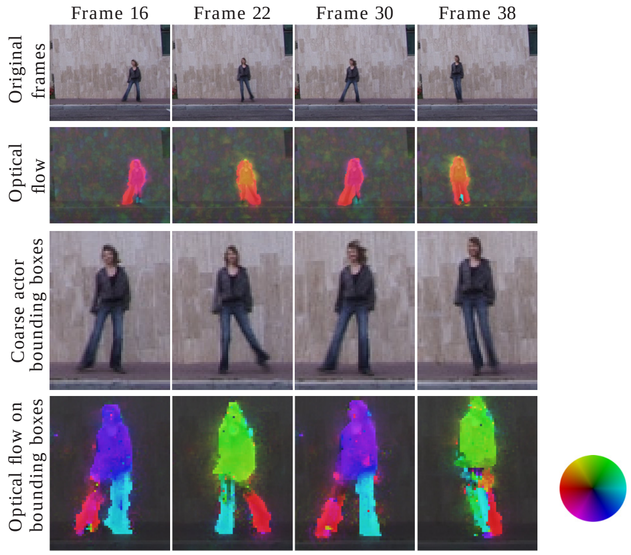


Fig. 3. Process of optical flow computation for 4 frames from Weizmann ‘`Side`’ action. Optical flow on whole frame essentially captures rigid body, translational motion, while optical flow on stacked, coarsely cropped, actor bounding boxes, successfully computes *articulated* motion. No pixel level segmentation is shown here.

patterns in surveillance videos. The details of our framework are described in the following subsections.

3.1 Low level Feature Computation

A common approach to obtaining motion is by quantizing it in terms of optical flow (e.g., HOF), computed from sequences of images depicting the action (as in [49], [13]). As noted in [10], it can be observed that whole-body translational motion is not helpful in discerning human actions, e.g., the difference between running and walking is due to the distinct patterns of articulation as compared to difference in speeds which is subjective and depends on factors such as camera location and orientation. On the other hand, computed optical flow in videos of such actions tends to be dominated by whole body translation, rather than the articulated motion (see Fig. 3-2$^{\text{nd}}$ row). It is therefore, desirable to compensate for such translational motion by approximate alignment of the agent performing the action, before computation of flow. To this end, we employ a simple process which includes computation of intensity difference images for consecutive frames, and thresholding of this difference image to obtain the motion blob, which is then represented as a coarse *bounding box*. These bounding boxes obtained in successive frames are then stacked together to obtain a sequence of cropped image frames for the entire training data set. The training is therefore performed in an unsupervised fashion, i.e., estimation of action primitives does not make use of the labels in the training data.

The simplicity of our proposed process can be compared with much stricter pre-processing assumptions of [36] (requiring body contours), [24] (which needs perfect foreground masks), [51] (landmark joint tracks), and [4] (employs HOG-based human detection instead of frame difference). Figure or actor centric spatiotemporal volumes were also required as input to the methods of Efros et al [9], and Fathi and Mori [10], who used human detection, background subtraction, or normalized cross-correlation based tracking to obtain such volumes.

The videos containing camera motion, for example, those in KTH or UCF Youtube datasets, are preprocessed to compensate for the moving camera, by feature based image alignment (homography from SIFT correspondences). It should be noticed that more complicated and comprehensive methods can be used in the pre-processing steps, e.g., Gaussian mixture model based background subtraction [38], human detection [5], and model based alignment of human blobs [29], etc.

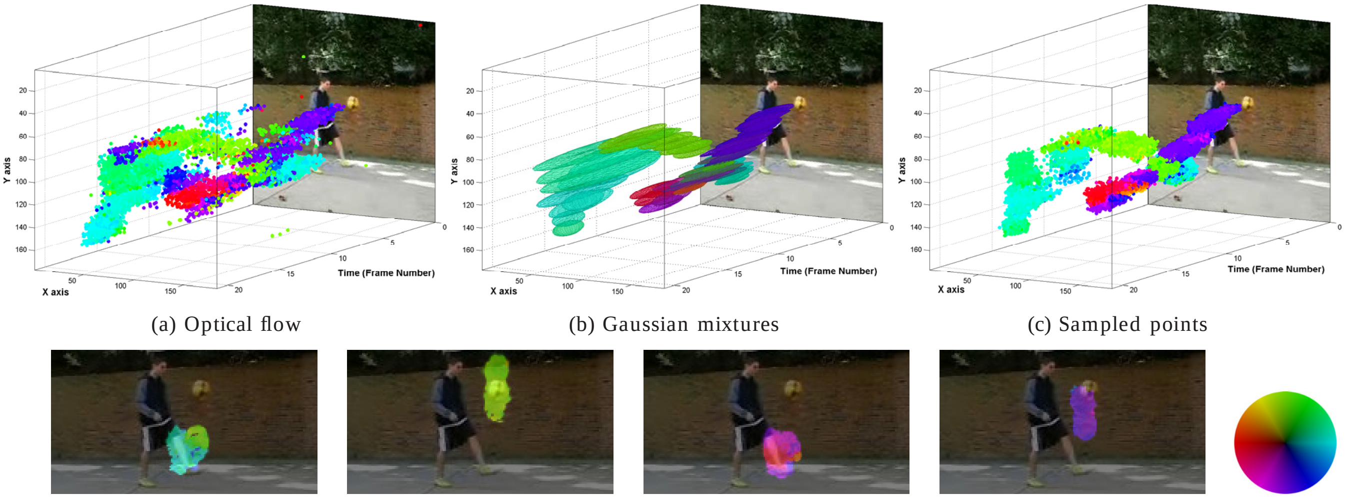
Lucas-Kanade optical flow is then computed for the *centralized* training videos. Some of the noise in optical flow is eliminated by removing flow vectors with magnitude below a small threshold. The resulting optical flow captures *articulated* motion as shown in Fig. 3. The flow vectors are then assumed to be values taken by the random variable, $\mathbf{f} = (x, y, u, v)$, where (x, y) is a location in the image, and (u, v) are the horizontal and vertical components of the optical flow vector at (x, y), as shown in Fig. 4(a). We propose that an action primitive be described as a Gaussian mixture distribution, $V_q = \{\mu_q, \mathbf{\Sigma}_q, \omega_q\}$, i.e.,

$$p_q(\mathbf{f}) = \sum_{j=1}^{N_q} \omega_j \mathcal{N}(\mathbf{f}|\mu_j, \mathbf{\Sigma}_j), \quad (1)$$

for the q^{th} primitive, where $1 \leq q \leq Q$, so that there are Q action primitives (or mixture distributions) in the entire dataset, which are referred to as $\mathbf{V} = \{V_q\}$. The goal of the training (or learning) phase then, is to estimate the parameters of each such mixture distribution, where the number of primitive actions (motion patterns), Q, as well as the number of components N_q, in each pattern's mixture distribution are unknown.

3.2 Gaussian Mixture Learning

We begin by performing a K-means clustering of all the 4d feature vectors obtained, as shown in Fig. 4(b). The value of K is not crucial and the goal is to obtain many, low variance clusters, which will become the Gaussian components in the motion patterns mixture distributions. In general, larger values of K will result in better performance, as shown later. It should be noted though, that the value of K does not affect the number of primitives obtained eventually, rather it controls the resolution or quality of the representation. The clustering is performed separately for D, short disjoint video clips, each of which contains k frames. The clustering results in a set of Z Gaussian components, $\mathbf{C} = \{C_z\}_{z=1}^{Z}$, for the entire training set, where, $Z = K \cdot D = N_1 + \ldots + N_Q$; and the mean, μ_z, covariance $\mathbf{\Sigma}_z$, and weight ω_z for the z^{th} component.



(a) Optical flow

(b) Gaussian mixtures

(c) Sampled points

left leg kicking

ball moving up

leg retraction

ball coming down

Fig. 4. Illustration of primitives discovery from optical flow: (a) optical flow for 20 frames from the UCF YouTube 's_juggling' action, shown as colored dots representing flow magnitude and orientation as brightness and color resp., as per color wheel. Several primitives corresponding to legs and ball motion are intuitively discernable. (b) K-means clustering and component linking results in statistical representation of primitive instances describing four sub-actions. Gaussian distributions are shown as error ellipses at 2.5σ, in $(x, y, magnitude)$, but placed in $(x, y, time)$. (c) optical flow data points *sampled* from the 4 mixture distributions are almost identical to the original data. Bottom row shows conditional expectation $E\left[\sqrt{u^2+v^2}, tan^{-1}(\frac{v}{u})|x, y\right]$, of flow for each of the 4 discovered sub-actions.

The eventual goal is to find out which of these components belong to each primitive action's distribution. We notice that the action primitive, represented as a motion pattern, repeats itself within the video of an action (because most action videos are cyclic), as well as within the training data set (because there are multiple examples of each action). Therefore, we first attempt to further group the Gaussian components, such that each repetition of a primitive is represented by such a high level group. We employ a Mahalanobis distance based measure to define a weighted graph, $G = \{\mathbf{C}, E, W\}$, where E and W are $Z \times Z$ matrices corresponding to edges and their weights. Whenever two components, C_i and C_j occur in consecutive, k-frames long, video clips, an edge exists between C_i and C_j, i.e., the element e_{ij} is 1. The weight for the edge between C_i and C_j is computed as sum of bidirectional squared Mahalanobis distances,

$$\begin{aligned} w_{ij} = & \left(\hat{\mathbf{f}}_i - \mu_j\right)^\top \Sigma_j^{-1} \left(\hat{\mathbf{f}}_i - \mu_j\right) \\ & + \left(\bar{\mathbf{f}}_j - \mu_i\right)^\top \Sigma_i^{-1} \left(\bar{\mathbf{f}}_j - \mu_i\right), \end{aligned} \tag{2}$$

where

$$\hat{\mathbf{f}}_i = (x_i + ku_i, y_i + kv_i, u_i, v_i) \tag{3}$$

is the forward transition prediction for the i^{th} component C_i, and

$$\bar{\mathbf{f}}_j = (x_j - ku_j, y_j - kv_j, u_j, v_j) \tag{4}$$

is the backward transition prediction for the j^{th} component. The two, 4d predicted features therefore serve as the points with respect to which the Mahalanobis distances are computed, essentially between pairs of Gaussian components. The weighted graph G, shown in Fig. 5, is then converted into an un-weighted one, by removing edges with weights below a certain threshold. The threshold is chosen as the Gaussian probability of a data point at 1.5σ ($\sim$87%), i.e., $w_{ij} \leq 2(1.5)^2 = 4.5$. A connected components analysis of this unweighted graph gives P sequences (mixtures) of Gaussian components, each of which is assumed to be a single occurrence of an action primitive, e.g., one instance of 'torso moving down'. Each such sequence of components (action primitive instance) is a Gaussian mixture, $S_p = \{C_m\}$, where $1 \leq p \leq P$, and $1 \leq m \leq M_i$, where M_i is the number of Gaussian components in the p^{th} mixture, S_p. We observe that these action primitives are shared in multiple similar actions, e.g., 'right arm going upwards' is shared by 'one hand waving' as well as 'both hands waving' (refer to Table 2 for more examples).

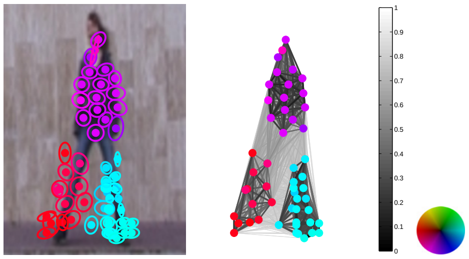
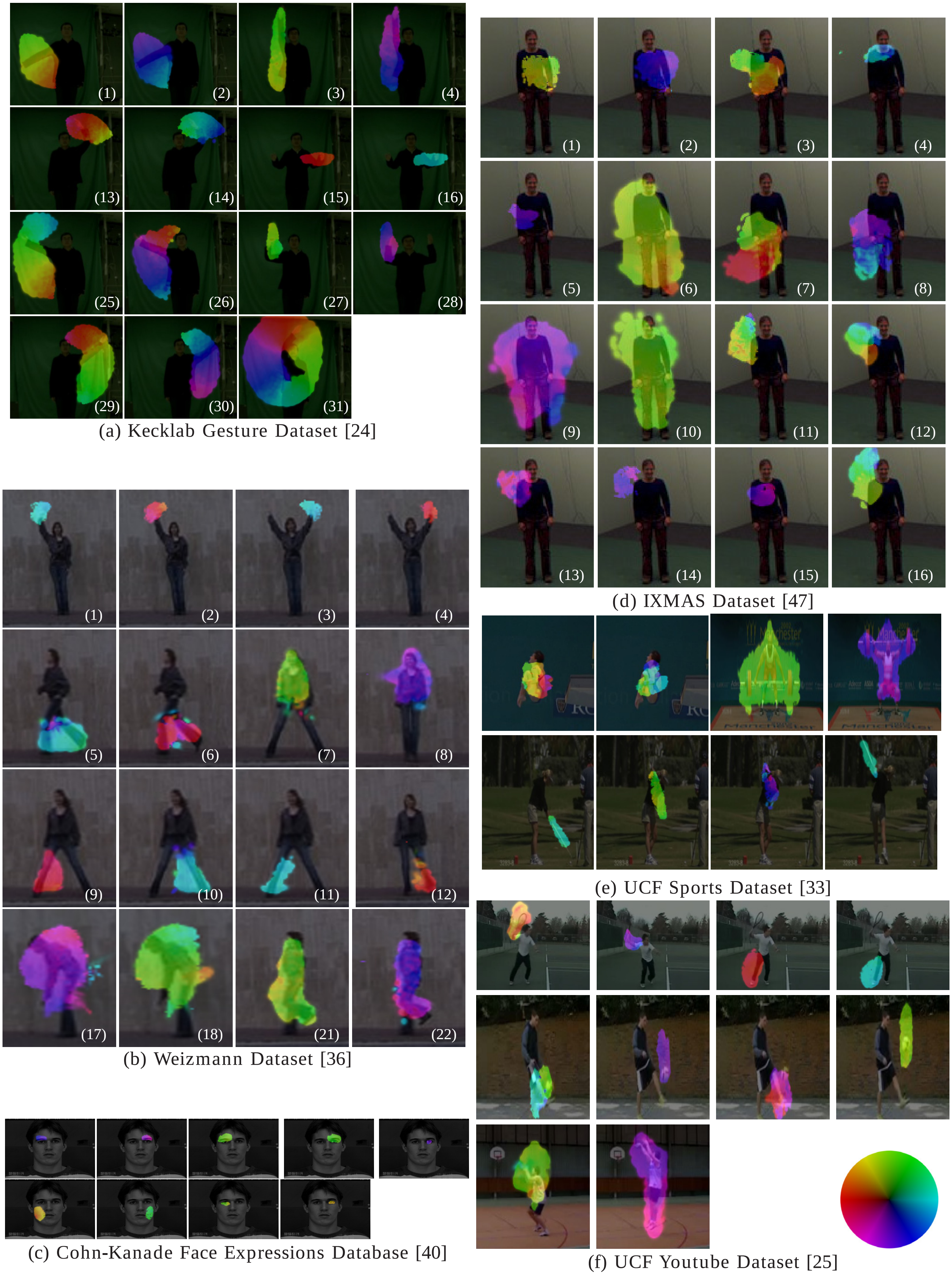


Fig. 5. Illustration of Graph G for components: (left) spatial means and covariances shown as colored dots and ellipses, with color corresponding to mean optical flow; (right) edge weights depicted by shades of gray.

As mentioned earlier, the instances computed are multiple occurrences of the same primitive actions. The final step in training is to merge the representations of these occurrences into a single Gaussian mixture for each primitive action, by computing the KL divergence



(a) Kecklab Gesture Dataset [24]

(d) IXMAS Dataset [47]

(b) Weizmann Dataset [36]

(e) UCF Sports Dataset [33]

(c) Cohn-Kanade Face Expressions Database [40]

(f) UCF Youtube Dataset [25]

Fig. 6. Some of the action primitives discovered in various datasets are shown by the conditional expected optical flow of the Gaussian mixtures that represent them. The direction is encoded by color as shown in the color wheel on bottom right, while the magnitude is depicted by brightness. See Table 2 for a list of actions represented by the primitives.

between each Gaussian mixture, and merging the ones with low divergences. The KL divergence between two mixtures, S_i and S_j is computed using Monte-Carlo sampling, and finally, we create a $P \times P$, non-symmetric positive matrix, Δ, of KL divergences between all P Gaussian mixtures of primitive instances, so that,

$$\Delta(i,j) = \mathcal{D}_{KL}(S_i \| S_j) = \sum_{n=1}^{N_{mc}} p_i(\mathbf{f}_n)\, log \left(\frac{p_i(\mathbf{f}_n)}{p_j(\mathbf{f}_n)}\right), \qquad (5)$$

where N_{mc} points are sampled from the Gaussian mixture S_i. A graph connected component analysis of the binarized Δ matrix then gives $\mathbf{V}$, the Q mixture models of the action primitives. Since true primitives occur multiple times in the dataset (due to multiple cycles and/or multiple example videos), primitives composed of less than 5 instances are removed as noise. Each action class is represented as a sequence or group of these primitives, which essentially define a vocabulary of the human actions. Examples of such Gaussian mixtures for different actions, gestures, and face expressions datasets are shown in Fig. 6 as conditional expected optical flow.

The expected values of optical flow magnitude and orientation for each pixel is a direct visualization of the 4d distribution of a motion pattern, which takes into account not only the flow at each pixel but also the probability of that pixel belonging to the motion pattern. The expected value of horizontal component of flow given a pixel location for the q^{th} motion primitive is computed as the weighted mean of conditional expectations of each component in the Gaussian mixture,

$$\mathbb{E}_q[\mathbf{u}|\mathbf{x},\mathbf{y}] = \sum_{j=1}^{N_q} \omega_j \mathbb{E}_j[\mathbf{u}|\mathbf{x},\mathbf{y}], \qquad (6)$$

and each component's expectation is given as,

$$\mathbb{E}_j[\mathbf{u}|\mathbf{x},\mathbf{y}] = \mu_u + \begin{bmatrix}\sigma_{ux}\\ \sigma_{uy}\end{bmatrix}^{\top} \begin{bmatrix}\sigma_{xx}\sigma_{xy}\\ \sigma_{yx}\sigma_{yy}\end{bmatrix}^{-1} \begin{bmatrix}x-\mu_x\\ y-\mu_y\end{bmatrix}, \qquad (7)$$

where the random variables $\mathbf{u}$ and $\mathbf{v}$ are assumed to be conditionally independent given location, and the random variables $\mathbf{x}$ and $\mathbf{y}$ take values (x,y) over the entire image (omitted in Eq. 7). The conditional expectation of the vertical component of flow, i.e., $\mathbb{E}_q[\mathbf{v}|\mathbf{x},\mathbf{y}]$, is computed in a similar manner. Therefore, Eq. 6 computes a scalar mean at a pixel, and we can finally obtain two 2d matrices, each as the conditional expected horizontal and vertical components of optical flow respectively.

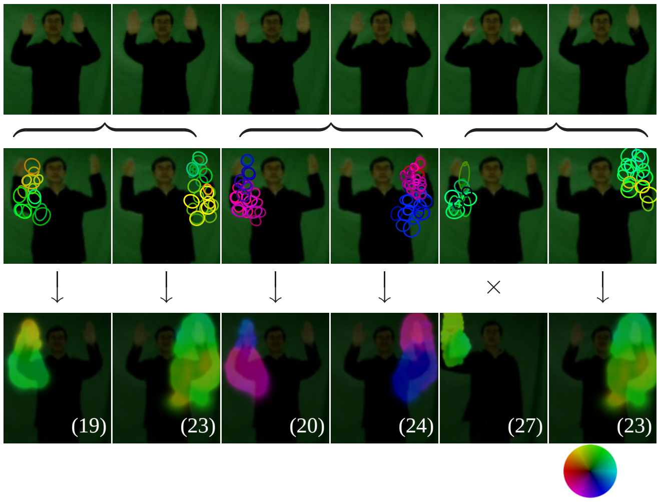


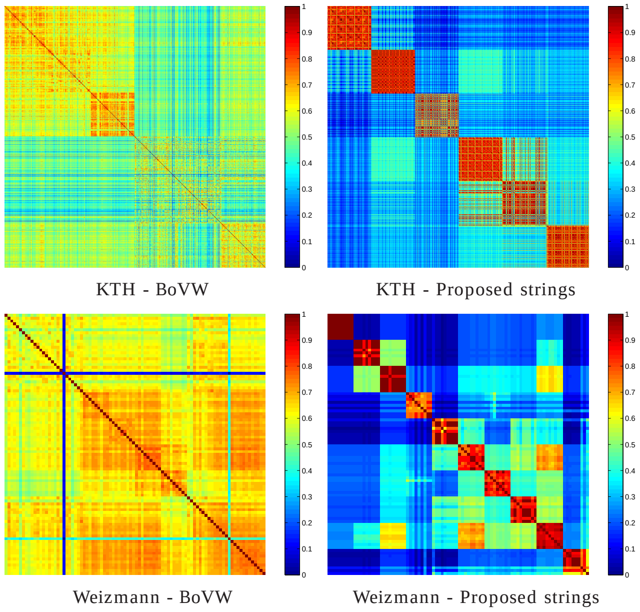
Fig. 7. Process of obtaining test video representation: (row 1): 6 frames (1.5 cycles) from Kecklab action '`go back`'. (row 2): 3 *pairs* of co-occurring primitive instances for the test video, shown as Gaussian error ellipses at 1.5σ (colors correspond to mean direction and brightness to mean magnitude). Horizontal braces ({) on top indicate co-occurring pairs. (row 3): results of primitive labeling using KL-divergence. Learned primitive with least divergence picked as label and shown at bottom. Downward arrows indicate correctness of labeling per primitive. The action *model* is represented by the sequence $T = (\{19,23\},\{20,24\})$. The only incorrect label is of the 5th detected primitive, labeled as 27 instead of 19. String matching score (to class model) for this video is 91.67%.

4 Action Representation

Given the automatic discovery, and statistical representation of action primitives, the next step in our proposed framework is to obtain a representation of the entire action video. We first deduce the occurrence of these primitives in a video. This process is straightforward for the videos used during primitive discovery, since we know which video each of the components in a Gaussian mixture came from. The i^{th} video is then represented as a temporal sequence of action primitive labels, i.e., $T_i = \{t_j\}$, where $t_j \in [1,Q]$.

For unseen test videos, this process is similar to the primitive discovery phase. However, since a test video is typically short, and contains at most a few cycles of the action, we do not perform the final step of primitive instance merging. This is because, for most test videos, only a single instance of action primitives is observed. We therefore obtain a set of motion primitives for a test video, and our goal is to relate these primitive instances to the previously learned representation, i.e., the action primitives learned from the entire training set which form the action vocabulary. This relationship is established by finding the KL divergence between each motion pattern observed in the test video, and all learned action primitives, and assigning it the label (or index) of the primitive with the least divergence. This process is illustrated in Fig. 7, where the second row shows patterns observed in a particular video, and the third row shows the corresponding primitives that each pattern was assigned to. The end result of this process then, is the representation of the given video, as a temporal sequence of action primitive labels, e.g., $T = (\{19,23\},\{20,24\},\{27,23\})$ in the example in Fig. 7.

We observe in our experiments, that most actions are adequately represented by very few primitives. This in part, is due to the nature of the primitive discovery



KTH - BoVW

KTH - Proposed strings

Weizmann - BoVW

Weizmann - Proposed strings

Fig. 8. Similarity matrices for *all* examples in two datasets using histogram intersection for BoVW, and string matching for *primitive strings*. Distinct classes are clearly discernable as squares of high similarity along the diagonal, for string matching matrices, lending themselves nicely to unsupervised clustering. The BoVW similarity matrices are largely random.

process and representation, where a specific sub-action of a limb, or the torso, is usually decomposed into at most a few primitives. The motion patterns thus represent the sub-actions at the highest possible level of abstraction, e.g., a continuous motion of the right arm moving upwards need not be broken further. Depending on complexity of articulation, an action can be represented by as few as one, or as many as ten or so primitives. Actual examples of strings representing different action classes are provided in Table 2.

For evaluation of the quality and discriminative nature of our proposed primitive actions, we put forth three different high level representations of an action video, all of which employ the observed primitives. These representations are described in detail in the following sections.

4.1 Histograms of Action Primitives

Given the primitive instances observed in an action video, the simplest way to represent the video is to ignore the temporal sequence, and collapse the sequence T into a histogram. The number of bins in this histogram is Q, the total number of primitives discovered in the dataset, and the value of each bin is the number of times that particular primitive is observed in the video. In order to make the histogram invariant to the number of action cycles, or the length of the video, the histogram is normalized to sum to 1. This histogram is analogous to the standard bag of video words histogram, and can be termed as a ‘bag of action primitives’ (BoAP), but has much fewer bins, is much more sparse, and therefore discriminative as compared to BoVW histograms. Specifically, given R instances discovered in the t^{th} test video, each a Gaussian mixture S_r, we generate an $R \times Q$ matrix, $\mathcal{D}_t$, where,

$$\mathcal{D}_t\left(r,q\right) = \frac{1}{\mathcal{D}_{KL}\left(S_r \| V_q\right)}, \tag{8}$$

and then create a Q dimensional, weighted histogram as,

$$\mathbf{H}_t = \sum_{r=1}^{R} \mathcal{D}_t\left(r,q\right). \tag{9}$$

Each bin in the histogram therefore, provides the likelihood of occurrence of a learned motion primitive in the test video. Given such a histogram, a supervised training based recognition (e.g., using nearest neighbor, or SVM classification), as well as, unsupervised action clustering (using histogram intersection as similarity measure) is performed and results are reported in Section 5.

4.2 Strings of Action Primitives

Another choice for action representation is to employ the string of primitive labels, T, instead of a histogram, which will preserve the temporal order of occurrence of each of the action primitives. The problem of action recognition then reduces to a simple string matching, where the letters in the string represent the primitive index between 1 and Q. Again, such strings can be used to perform supervised nearest neighbor classification, as well as unsupervised clustering using string matching scores as similarity measure.

We perform string matching using the well known Needleman-Wunsch algorithm [28], which is a linear global optimization method. We therefore obtain a confidence score in a particular matching. The matching score between two action videos, i and j, is written as $\Theta(T_i, T_j)$, and the scores for all possible alignments are computed using the following recursive relationship:

$$\begin{aligned} A(m,n) &= max \begin{cases} A(m-1,n-1) + b(T_i^m, T_j^n) \\ A(m-1,n) - g \\ A(m,n-1) - g \end{cases} \\ A(m,0) &= -mg \\ A(0,n) &= -ng \\ A(0,0) &= 0, \end{aligned} \tag{10}$$

where,

$$b\left(T_i^m, T_j^n\right) = \begin{cases} 1 : T_i^m \equiv T_j^n, \\ -0.5 : otherwise \end{cases} \tag{11}$$

and g is the gap penalty set to -1. The matching score $\Theta(T_i, T_j)$ is the maximum score alignment, $A(m,n)$.

For each action class, we have as many strings as the number of examples. A test string is compared to each training string, and is assigned the label of the class with the highest matching score, essentially a nearest neighbor approach. In order to handle co-occurring pairs (or groups), they are sorted arbitrarily, and for string matching, all possible permutations are tried, and the best chosen. Despite the simplicity of the recognition approach, very encouraging results are obtained as reported in Section 5, which is a direct consequence of the discriminative nature and high level of abstraction of the proposed action representation. In actuality, we observed that even a literal string matching performs reasonably well without global alignment [28] of primitive labels. In order to test if the motion patterns based representation is discriminative enough, we visualize histogram intersection and string matching based similarities between examples of actions, some of which are shown in Fig. 8.

4.3 Temporal Model of Action Primitives

Given that in our framework a video is essentially a temporal sequence of primitive actions, where each primitive is observed with a specific probability or confidence (KL divergence), one of the most succinct and information preserving ways to represent an action class is by learning a Hidden Markov Model (HMM) for the primitive sequences. HMMs [31] have been used extensively in speech recognition, object tracking, as well as human action recognition. Given discrete, finite valued time series of action primitives, T_i as representation of the i^{th} video, our goal is to learn a discrete HMM, $\lambda_l = \{A_l, B_l, \pi_l\}$, for each action class $l \in [1, \mathcal{L}]$, where A_l is the state transition matrix, π_l is the initial distribution or prior, and B_l represents the probability distributions for the observed feature vector conditional on the hidden states, which are assumed to be represented as a Gaussian mixture model as is traditionally done. The maximum likelihood approach is then used to classify each action example:

$$l^* = \operatorname*{argmax}_{l} \; P(T_i|\lambda_l), \tag{12}$$

that is, the conditional probability of an action video i, represented by the feature vector T_i, the sequence of action primitive labels, given the model for action l^* is maximum among all classes. The number of states for each class model was chosen to be 5 in all our experiments.

5 Experiments

The proposed primitive representation has been evaluated for five human action datasets, as well as a composite dataset, two human gestures datasets, and a facial expressions database. We tested our representation using three different high level representations (strings, histograms, and HMMs), for three distinct applications (unsupervised clustering, 1/k-shot recognition, and conventional supervised recognition), to show representation and recognition quality and performance. Our extensive experiments on a variety of datasets provide insight into not only how our framework compares with state-of-the-art, but also into the very nature of the action recognition problem. For most of the experiments, we demonstrate the superiority of the proposed representation compared to existing methods as detailed in the following subsections. Some key statistics related to learning of action primitives as motion patterns, are summarized in Table 1.

TABLE 1
Some statistics related to experiments. Notice that as few as 55 action primitives sufficiently represent even large datasets (e.g., composite dataset with 25 classes).

Supervised	Value of K	# primitive instances	# initial primitives	# final primitives
Kecklab [24]	20	620	45	31
Weizmann [36]	30	2000	52	35
KTH [35]	50	15000	60	29
UCF Sports [33]	50	3500	400	300
UCF YouTube [25]	50	4000	400	200
Cohn-Kanade [40]	30	740	41	24
Unsupervised				
Kecklab [24]	50	700	50	37
Weizmann [36]	50	2500	43	26
KTH [35]	50	15500	51	20
IXMAS [47]	50	820	48	30
Composite	50	19800	88	55
UCF Sports [33]	50	3500	450	211
UCF YouTube [25]	50	4200	490	309
Cohn-Kanade [40]	30	740	56	37

TABLE 2
Action primitives used in each of Kecklab gesture, and Weizmann datasets to represent different action classes. A primitive can be part of multiple actions, while an action may be represented by a single primitive. Primitives grouped by {} co-occur in a clip.

Kecklab Gesture

Action	Primitives	Action	Primitives
Turn left	1, 2	Attention right	13, 14
Turn right	5, 6	Attention both	{9, 14}, {10, 13}
Turn both	{1, 5}, {2, 6}	Speed up	27, 28
Stop left	3, 4	Come near	{19, 23}, {20, 24}
Stop right	7, 8	Go back	{17, 21}, {18, 22}
Stop both	{3, 7}, {4, 8}	Close distance	{11, 16}, {12, 15}
Attention left	9, 10	Start engine	31

Weizmann

Action	Primitives	Action	Primitives
Bending	17, 18	Sideways gallop	7, {9, 10}, 8, {11, 12}
1 hand Waving	1, 2	Walking	{13, 14}, {15, 16}, 19, 20
2 hand waving	{1, 4}, {2, 3}	Jumping	23, 24, 25, 7, 8
Running	5, 6	Jumping Jacks	7, {9, 10}, {1, 4}, 8, {11, 12}, {2, 3}
Para-jumping	7, 8	Skipping	21, 22

5.1 Unsupervised Clustering

As mentioned earlier the problem of collecting a large number of labelled examples for training is a laborious task. On the other hand, in practical scenarios the available videos to be recognized or classified are mostly unlabelled. Indeed the vast amount of visual data in the public domain falls in this category, e.g., web sources like YouTube, etc. It is therefore desirable to attempt group-

TABLE 3
Quantitative comparison of different representations for unsupervised clustering with and without actor centralization. BBx in column 2 implies ‘bounding box’.

Method	Bag of video words				
	No BBx Not Dense	Dense Sampling in Bounding Box			Action Primitives
Dataset	Dollar [8]	Dollar [8]	ISA [23]	MBH [43]	(proposed)
Weizmann	67%	69.2%	69.9%	72.5%	91%
KTH	65%	61.0%	63.6%	67.2%	91%
IXMAS	53%	51.2%	53.6%	67.2%	63%
UCF Sports	63%	59%	61.7%	61.9%	68%
UCF Youtube	49%	37%	36.7%	41.2%	54%
Composite	43%	41%	42%	45.8%	79%

ing of such videos into meaningful categories, without provision of training examples or user intervention. This problem is referred to as unsupervised clustering.

In this experiment, *all* videos in the dataset are used to learn the action primitives representation, and the videos are represented as strings of primitive labels. A value of 50 was used for K (in k-means) for all datasets except the Cohn-Kanade face expressions databased, where $K =$30. A string matching similarity matrix of all videos is then constructed (Fig. 8), and clustering is performed by thresholding and graph connected components to obtain groups of videos. Each video in a cluster is then assigned the same label as that of the dominating class within the cluster, and comparison of the assigned label with the ground truth label, provides classification accuracy. The results of these experiments on most datasets are summarized in Table 3.

This experiment truly reveals the discriminative power of any representation, because if the representation is truly unique, one would expect a high intra-class similarity and a high inter-class distance in the feature space, thus reasonably lending the features or data points to clustering. In order to simplify interpretation of results, we fixed the number of clusters to the number of classes in the dataset. It can be observed that the same experiment can be performed without this constraint. The labels of the videos however, were not used during the entire process (except evaluation). Obviously, this experimental setup will achieve much lower classification scores as compared to supervised recognition, but our goal is to compare the results across different representations, including the ones achieving state-of-the-art performance in traditional supervised recognition scenarios.

As reported in Table 3, the proposed action primitives based representation outperforms all other methods for nearly all datasets. One can argue (and rightly so) that a comparison of unsupervised clustering using the proposed algorithm is not comparable to other techniques due to the advantage of actor centralization. We therefore compared a number of existing techniques using the same exact actor bounding boxes for all methods. Moreover, since our action primitives can be interpreted as dense, pixel-level representation, we performed the quantitative comparison using other dense features as well. We can make several interesting observations from the results presented in Table 3.

First, among some of the best existing features, the motion boundary histogram (MBH) feature [43], consistently performs the best.

Second, as expected, the use of actor centralized volumes instead of features computed on the full frame performs comparably in relatively simpler videos, but as they becomes more complex this trend is reversed. The reason for this result is that most of the ‘important’ or discriminative features (video words) tend to appear on the background instead of the actor’s body. Therefore, for ‘actions in the wild’ videos, the success of bag of words based methods mostly relies on capturing the scene information. The same observation has been made by Kuehne et al [21], where it is shown that the Gist descriptor performs only 5% worse than HOG / HOF on UCF YouTube, and actually performs 2% better than HOG / HOF on UCF Sports. Features capturing the background scene information are indeed useful in improving quantitative performance, but they obviously are not representative of the action (motion) itself, rather the characteristics of the particular dataset. Indeed, as we report later, by augmenting our representation with scene information, we were able to reach close to the state-of-the-art in supervised recognition.

Third, we observe that the proposed action primitives based representation outperforms all other methods on all datasets with very significant margins, with the exception of IXMAS dataset, where dense MBH features on actor bounding boxes performs $\sim$4% better.

- **Composite Dataset**: We also experimentally demonstrate and quantify our claim that the proposed representation is flexible and discriminative. In addition to recognition from very few examples, and achieving significantly better accuracy for unsupervised clustering compared to conventional methods, another test of these claims is to perform action recognition across multiple datasets, i.e., truly representative features, vocabulary, or primitives of an action should be common for a class, across datasets.

We perform this experiment by combining three sources, namely, the KTH, IXMAS, and the Weizmann datasets to obtain a ‘Composite’ set of 25 action classes. We performed a variety of experiments with this dataset. First, the classification accuracy for the composite dataset is obtained by attempting unsupervised (unlabeled) clustering of videos. The action primitives are learned using *all* the videos of the composite set, and as shown in Table 1, 55 primitive actions are discovered. As listed in Table 3, clustering by thresholding a graph of string similarity scores resulted in a classification accuracy of **79%**, compared to only 43% for BoVW using a codebook of size 500 (>9 times the size of our vocabulary). The improvement of 36% over 25 action categories, even with a more compact vocabulary, is only possible if most of the learned action primitives are exclusive to specific action classes, thus resulting in representations of videos that are sparse in terms of the primitives. Our approach significantly outperforms the state-of-the-art descriptors of independent subspace analysis (ISA) [23] and motion boundary histograms (MBH) [43], even when all methods employ the same figure-centric volumes. The confusion table obtained by classification via clustering is shown in Fig. 9(b).

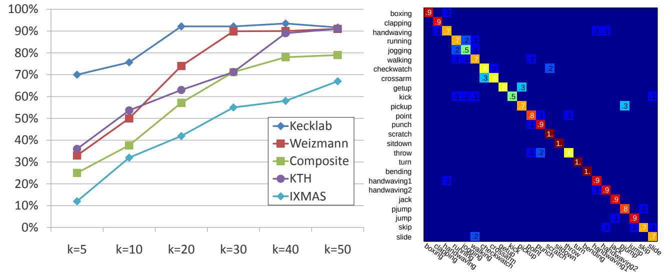
- **Cross-dataset Grouping**: Another interesting experiment performed using the Composite dataset was, to use the 55 action primitives learned in the *Composite* dataset (Weizmann + KTH + IXMAS), to represent and classify videos in the 3 datasets individually, and compare the performance against the representation using dedicated, indigenous primitives learned within each dataset (i.e., against results reported in Table 3). Naturally, it is expected that the performance would degrade using the composite primitives, due to larger variance in shared actions, and more noise. However, a meaningful high-level representation should be largely robust to these problems, which is what we observed in our experiments. The performance using composite primitives on Weizmann, KTH, and IXMAS, was **86%**, **88%**, and **59%** respectively, compared to 91%, 91%, and 63% respectively for the individual datasets, which is a very insignificant deterioration.

- **Kecklab Gesture Dataset**: Unsupervised clustering over the entire Kecklab gesture dataset was also performed using the proposed action primitives representation, as described before, and the proposed method obtained an average classification accuracy of **91.64%**. This performance is very close to supervised recognition accuracy using labeled training examples with the same representation, i.e., 94.64% (Table 7).

- **Cohn-Kanade Expressions Database**: The application of unsupervised clustering on the facial expressions database resulted in an average accuracy of **82.1%** which is comparable to supervised learning accuracy of 81.0% ([33]) and 86.7% (proposed representation).

- **Effect of Parameter *K***: We also quantified the effect of the parameter, K, the number of components in each video clip obtained using the K-means algorithm. The larger values of K essentially correspond to increased granularity of representation (even though the distribution is defined continuously space and flow space). We observed that as conjectured earlier, a high, computationally reasonable value of K, lets the performance of our method peek and level out. This can be observed in Fig. 9(a) where results of unsupervised clustering are quantified for different values of K.

As mentioned earlier, it should be noted that although the performance of the proposed approach as well as competitive methods for unsupervised clustering is lower than the state of the art, it is a much harder problem as well. This is due to the lack of labeled training examples, which can be used to learn classification boundaries between positive and negative examples, even when the data points representing videos cannot otherwise be grouped in the high dimensional space. To conclude this section, we summarize two main points: (1) a discriminative representation is one where visual similarity (in feature space) is highly correlated with semantic similarity (same class/category), and should therefore allow feature based grouping of unlabelled videos; and (2) in order to be applicable to real life, practical scenarios, the representation of an action should capture the action (motion and articulation) of the actor (albeit with some static context), rather than the background scene which may serve to artificially inflate performance.



(a) (b)

Fig. 9. Classification accuracy by unsupervised clustering using proposed primitives. (a) shows the increase in performance as the value of K, in the K-means clustering is increased. The confusion matrix for the 25 class 'composite' dataset is shown in (b); avg. accuracy was **79%**.

5.2 One-shot and K-shot Learning

To quantify the discriminative power of our representation, we attempted action recognition using as few as possible training examples using the proposed method. This experiment was performed for Kecklab Gesture, Weizmann, and UCF YouTube datasets, as well as Cohn-Kanade face expressions database, and the recently posed ChaLearn Gesture Challenge dataset [1].

Fig. 10 shows the performance of the proposed representation using a variable number of training examples, as well as comparison to BoVW framework with same settings using Dollar [8] features. For the Kecklab training dataset, 9 examples per action class are available, and we performed incremental learning and recognition increasing the number of available videos from 1 to 9, while testing on the entire test set in each increment. For the Weizmann dataset, we trained using an increasing number of videos (1 to 8), as we added videos from each of the 9 performers incrementally. In each increment, all the unused videos were tested upon. Therefore, in the first increment, videos from 8 actors were tested using videos from 1 actor as training, and vice versa for the last increment. For both datasets, using even a *single* training

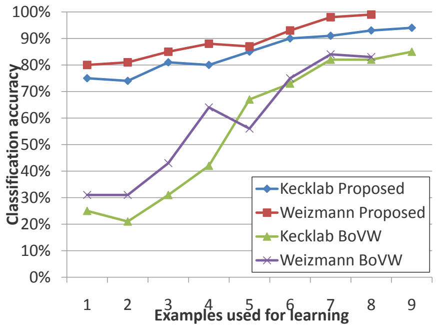


Fig. 10. Classification of actions using a variable number of training examples. The values corresponding to 1 on the X-axis are essentially results of *one-shot learning*. Using our method, an average accuracy of around 80% is obtained for both the Kecklab gesture and Weizmann datasets, using a *single* training example per class for primitive discovery and training video representation. The proposed method outperforms BoVW by a large margin.

TABLE 4
Quantitative comparison of different methods for one-shot learning on ChaLearn dataset [1] (average accuracy for development batches 1-10)

Method	Average accuracy
ISA [23]	25.4%
MACH 3D [33]	34.67%
MBH [43]	32.4%
STIP [22]	16.4%
Proposed action primitives	56.34%

example per action class, around 80% recognition rate is achieved, compared to about 30% for BoVW.

- **ChaLearn Gesture Challenge Dataset**: Recently a new comprehensive dataset of videos of human actors performing a variety of gestures has been made available to researchers under the Microsoft ChaLearn Gesture Challenge [1]. The goal of the challenge is to employ systems to perform gesture recognition from videos containing diverse backgrounds, using a single example per gesture, i.e., one-shot learning.

We have used this dataset to test and compare the ability of our proposed representation for one-shot learning. Specifically, we used the first 10 development batches out of the available hundreds. Each batch has approximately 15 training and 20 test videos. The videos are captured from frontal views with the actor roughly centralized and no camera motion. The actor centralization step was not performed for this dataset. Although the gestures performed in this dataset were simultaneously captured from a color video camera as well as a depth camera (using the Kinect^{TM} camera), we only used the RGB videos for our experiments. Table 4 shows a comparison of different well-known approaches to the proposed technique. It should be noted that all experimental settings for these comparative evaluations were the same. As evident from the quantitative comparison, the proposed representation is well suited to capturing human motion from a single example, and is by far the best performer.

TABLE 5
Results of k-shot recognition on the Cohn-Kanade face expression database. BoW framework employs MBH [43] with actor centralization and dense sampling.

# of examples	1	4	7	10
BoW	41.2%	44.5%	52.1%	60.5%
Proposed	65.1%	70.9%	74.2%	79.3%

TABLE 6
Comparison of k-shot recognition on UCF YouTube dataset. BoW framework employs MBH [43] with actor centralization and dense sampling.

# of examples	1	10	20	30	40	50
BoW	11.2%	20.6%	28.9%	35.4%	39%	43%
Proposed	19.3%	31.3%	39.2%	46.3%	50%	51%

We also compared the performance of our approach against the state-of-the-art MBH descriptor [43] using the same settings as action primitives (actor centralization and dense sampling), for the Cohn-Kanade face expressions and UCF YouTube datasets. The results of these comparisons are shown in Tables 5 and 6. For the Cohn-Kanade database, we randomly chose 1, 4, 7, and 10 examples for training, and tested on the rest. The results were averaged over 10 runs. Similarly for the UCF YouTube dataset, 1, 10, 20, 30, 40, and 50 videos for each class, and tested on the rest. The reported accuracies were averages of 50 runs. For all the experiments, the BoVW method used exactly the same training and testing samples. The only difference is that the proposed approach learns action primitives, while the BoVW framework learns the codebook from the training samples. Given the same advantage of figure-centric volumes to both approaches, action primitives demonstrate their highly discriminative yet flexible nature in one and k shot recognition.

5.3 Supervised Recognition

Finally, for the sake of completeness and comparison to results in the existing literature, we present our results using the traditional supervised learning approach. In this experiment, a dataset is divided into training and testing sets. The primitive action representation is learned from examples in the training sets. The training videos are then represented as strings of the primitive labels. Given a test video, pattern instances are estimated using the proposed approach, which are then represented as Gaussian mixtures. These distributions are then compared against the learned primitive distributions using KL divergence, and labeled. The test video is thus also represented as a string of learned primitives. Finally, a string matching based nearest neighbor classifier is employed to assign an action label to the test video. The results on different datasets using this approach are reported in Table 7.

The experimental settings (train-test partitions, cross validation methodology, etc.) used in our experiments are the same as the original papers. The Kecklab gesture dataset [24] is already partitioned into training and test sets. The Weizmann dataset [36], which consists of videos of 10 actions performed by 9 individual actors, is tested using leave-one-out cross validation as in [36]. UCF Sports and YouTube datasets were experimented with using the original settings of the corresponding papers, i.e., [33] (leave-one-out) and [25] (25 fold cross validation) respectively. 25 fold cross validation was also performed for KTH.

In order to provide a fair comparison, the first 3 rows of Table 7 use actor centralization and dense sampling of features within actor bounding boxes (same as action primitives). Moreover, instead of a string matching based nearest neighbor classification, we use a histogram of action primitive labels trained using an SVM classifier (same as ISA and MBH). For string matching and HMM based nearest neighbor recognition, the UCF Sports dataset was recognized with average accuracies of 61% and 85% respectively. These accuracies for the UCF YouTube dataset were 42% and 51% respectively. The worse performance of strings and HMM can be attributed to the fact that temporal order is not too helpful within short videos where actions are represented by very few primitives.

On simpler datasets like Kecklab gesture, Weizmann, and KTH, it can be observed that the performance of using the actor centralized videos is almost the same as the state-of-the-art features using original full frames. However, on more complex videos like the UCF Youtube dataset, a significant drop in performance from the original videos to centralized videos is noticeable for not only the proposed approach but also state-of-the-art descriptors like ISA and MBH. As mentioned earlier in Section 5.1, the reason for this drop is that many of the video words contributing to discriminative power of the histogram feature, appear on the background scene instead of the human body. The goal of our action representation framework however, is not to capture static scene properties, but the motion and articulation of the actor. To verify our hypothesis about the reason for this performance drop, we augmented the action primitives histogram with dense SIFT bag of words feature computed on only the non-actor pixels, and were able to improve the performance from 57% to 79.5%, and to 86.9% when we used MBH features instead of SIFT. So our method performed comparably to the state-of-the-art when recognition is truly based on the actor's motion.

• **Cohn-Kanade Expressions Database**: Our representation is also readily applicable to subtle motions like videos of facial expressions. We tested our hypothesis on the Cohn-Kanade AU-Coded Facial Expression Database [40], which contains videos of multiple humans depicting various expressions corresponding to human emotions. Facial action units (AU) have often been used to describe expressions. Our experiments were carried out to classify action units into one of seven upper face AU classes, as was done in [33]. Using 24 motion primitives (some of which are shown in Fig. 6), the proposed method achieved an average accuracy of **86.7%** using four fold cross validation, compared to 81% in [33].

6 Conclusion

This paper has proposed a method that automatically discovers a flexible and meaningful vocabulary of actions using raw optical flow, learns statistical distributions of these primitives, and because of the discriminative nature of the primitives, very competitive results are obtained using the simplest recognition and classification schemes. Our representation offers benefits like recognition of unseen composite action, insensitivity to occlusions (partial primitive list), invariance to splitting of primitive during learning, detection of cycle extents and number, etc. The meaningful nature of the primitives is also promising towards closing the gap between visual and textual representations and interpretations.

References

[1] http://gesture.chalearn.org/.

[2] M. Ahmad and S. Lee. Human action recognition using shape and CLG-motion flow from multi-view image sequences. *Pattern Recogn.*, 41:2237–2252, July 2008.

[3] A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *WACV*, 1996.

[4] C. Chen and J. Aggarwal. Recognizing human action from a far field of view. In *WMVC*, 2009.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] T. Darrell and A. Pentland. Space-time gestures. In *CVPR*, 1993.

[7] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *CVPR*, 1997.

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, 2005.

[9] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[10] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008.

[11] R. Filipovych and E. Ribeiro. Learning human motion models from unsegmented videos. In *CVPR*, 2008.

[12] A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *ECCV*, 2008.

[13] J. Hoey and J. Little. Representation and recognition of complex human motion. In *CVPR*, 2000.

[14] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010.

[15] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.

[16] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005.

[17] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007.

[18] A. Kläser, M. Marszałek, I. Laptev, and C. Schmid. Will person detection help bag-of-features action recognition? Technical Report RR-7373, INRIA Grenoble - Rhône-Alpes, 2010.

[19] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.

[20] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009.

[21] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.

[22] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

TABLE 7
Quantitative comparison of the proposed action primitives with some of the state-of-the-art techniques for supervised action recognition. Note that not all of these methods employ the same experimental settings. Nevertheless, these statistics provide a bird's eye view of where our framework stands with respect to the related work, despite its simplicity. The first 3 rows however, use the same settings, i.e., actor centralization, and dense sampling.

Kecklab	Accuracy	Weizmann	Accuracy	KTH	Accuracy	UCF Sports	Accuracy	UCF YouTube	Accuracy
Proposed	94.6	Proposed	99	Proposed	93.2	Proposed	88	Proposed	57
ISA [23]	93.2%	ISA [23]	91.7%	ISA [23]	-	ISA [23]	82.5%	ISA [23]	51.2%
MBH [43]	91%	MBH [43]	96.3%	MBH [43]	-	MBH [43]	85.6%	MBH [43]	55.6%
Lin [24]	91	Wang [46]	97.2	Liu [26]	94.1	Rodriguez [33]	69.2	Liu [25]	71.2
		Jhuang [15]	98.8	Jhuang [15]	91.6	Wang [44]	85.6	Ikizler-Cinbis [14]	75.21
		Thurau [39]	94.4	Lin [24]	95.7	Kovashka [19]	87.27	Wang [43]	84.2
		Lin [24]	100	Ahmad [2]	87.6	Kläser [18]	86.7		
		Fathi [10]	100	Wang [46]	92.4	Wang [43]	88.2		

[23] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011.

[24] Z. Lin, Z. Jiang, and L. Davis. Recognizing actions by shape-motion prototype trees. In *ICCV*, 2009.

[25] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *CVPR*, 2009.

[26] J. Liu and M. Shah. Learning human action via information maximization. In *CVPR*, 2008.

[27] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004.

[28] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, March 1970.

[29] O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. In *CVPR*, 2010.

[30] V. Parameswaran and R. Chellappa. View invariants for human action recognition. In *CVPR*, 2003.

[31] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, 1989.

[32] N. Robertson and I. Reid. Behaviour understanding in video: a combined method. In *ICCV*, 2005.

[33] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[34] I. Saleemi, L. Hartung, and M. Shah. Scene understanding by statistical modeling of motion patterns. In *CVPR*, 2010.

[35] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

[36] E. Shechtman, L. Gorelick, M. Blank, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 29(12):2247–2253, 2007.

[37] V. Singh and R. Nevatia. Action recognition in cluttered dynamic scenes using pose-specific part models. In *ICCV*, 2011.

[38] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22(8):747–757, 2000.

[39] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008.

[40] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *PAMI*, 23:97–115, 1999.

[41] K. Tran, I. Kakadiaris, and S. Shah. Modeling motion of body parts for action recognition. In *BMVC*, 2011.

[42] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine Recognition of Human Activities: A Survey. *IEEE T-CSVT*, 18(11):1473–1488, 2008.

[43] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[44] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[45] X. Wang, X. Ma, and G. Grimson. Unsupervised activity perception by hierarchical Bayesian models. In *CVPR*, 2007.

[46] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *ICCV HMUMCA Workshop*, 2007.

[47] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, 2007.

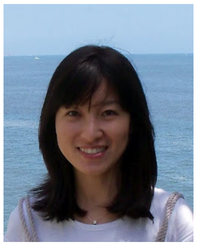
[48] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *CVIU*, 115(2):224–241, 2010.

[49] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. In *ICCV*, 1998.

[50] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *CVPR*, 2005.

[51] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *ICCV*, 2005.

[52] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, 2001.



Yang Yang received her B.S. degree in electrical engineering from Beijing University of Technology, Beijing, China, in 2008. She is currently working towards her PhD degree in the Computer Vision Laboratory, University of Central Florida. Her research interests include action and event recognition, scene understanding, manifold learning and deep learning.

Imran Saleemi is a Postdoctoral Associate at the Computer Vision Laboratory at University of Central Florida. His research in computer vision is in the areas of visual tracking, multi-camera and airborne surveillance, statistical modeling and estimation of motion patterns, and probabilistic representation of actions and complex events. He received the BS degree in Computer System Engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan in 2004, and the MS and PhD degrees, both in Computer Science, from the University of Central Florida in 2008 and 2011 respectively.

Mubarak Shah , Agere Chair Professor of Computer Science, is the founding director of the Computer Vision Lab at the University of Central Florida. He is a co-author of three books and has published extensively on topics related to visual surveillance, tracking, human activity and action recognition, object detection and categorization, shape from shading, geo registration, and visual crowd analysis. Dr. Shah is a fellow of IEEE, IAPR, AAAS and SPIE. He is an ACM Distinguished Speaker. He was an IEEE Distinguished Visitor speaker for 1997-2000, and received IEEE Outstanding Engineering Educator Award in 1997. He is an editor of international book series on Video Computing; editor in chief of Machine Vision and Applications journal, and an associate editor of ACM Computing Surveys journal. He was an associate editor of the IEEE Transactions on PAMI and the program co-chair of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.