A Multi-Scale Hierarchical Codebook Method for Human Action Recognition in Videos Using a Single Example

Mehrsan Javan Roshtkhari, Martin D. Levine
Centre For Intelligent Machines, Department of Electrical and Computer Engineering
McGill University, Montreal, QC, Canada
Email: {javan,levine}@cim.mcgill.ca

***Abstract*—This paper presents a novel action matching method based on a hierarchical codebook of local spatio-temporal video volumes (STVs). Given a single example of an activity as a query video, the proposed method finds similar videos to the query in a video dataset. It is based on the bag of video words (BOV) representation and does not require prior knowledge about actions, background subtraction, motion estimation or tracking. It is also robust to spatial and temporal scale changes, as well as some deformations. The hierarchical algorithm yields a compact subset of salient codewords of STVs for the query video, and then the likelihood of similarity between the query video and all STVs in the target video is measured using a probabilistic inference mechanism. This hierarchy is achieved by initially constructing a codebook of STVs, while considering the uncertainty in the codebook construction, which is always ignored in current versions of the BOV approach. At the second level of the hierarchy, a large contextual region containing many STVs (Ensemble of STVs) is considered in order to construct a probabilistic model of STVs and their spatio-temporal compositions. At the third level of the hierarchy a codebook is formed for the ensembles of STVs based on their contextual similarities. The latter are the proposed labels (codewords) for the actions being exhibited in the video. Finally, at the highest level of the hierarchy, the salient labels for the actions are selected by analyzing the high level codewords assigned to each image pixel as a function of time. The algorithm was applied to three available video datasets for action recognition with different complexities (KTH, Weizmann, and MSR II) and the results were superior to other approaches, especially in the cases of a single training example and cross-dataset action recognition.**

***Keywords*-action recognition; bag of video words; hierarchical codebook.**

I. Introduction

Given the tremendous number of potential practical video applications, there is a great demand for automated systems that analyze and understand the contents of these videos. Obviously, recognizing and localizing human actions in a video are of primary importance to such a system. To date, in the computer vision community, "action" has largely been taken to be a human motion performed by a single person, *lasting for just a few video frames*, taking up to a few seconds, and containing one or more events. Walking, jogging, jumping, running, hand waving, picking up something from the ground, and swimming are some examples of such human actions [1]. In this paper, our main goal is to address the problem of *action recognition* in real environments using a hierarchical probabilistic video-to-video matching framework. To achieve this, we have developed a fast data-driven approach which finds similar videos to a single labeled "query" video in a "target" set. Assuming that the query contains an action of interest, e.g., walking, we find all videos in the target set that that are similar to the query, i.e., those that contain the same activity. This video-to-video comparison also makes it possible to label activities, the so-called action classification problem. Our approach does not require long video training sequences, object segmentation, tracking or background subtraction. The method can be considered as an extension to the original *Bag of Video Words* (BOV) approach for action recognition. An overview of the algorithm is presented in Figure1.

Although the initial spatio-temporal volumetric representation of human activity eliminates some pre-processing steps, such as background subtraction and tracking, it does share some of the common drawbacks with methods that do require these. For example, in general BOV-based approaches for activity recognition involve salient point detection and are also unable to handle scale variations (spatial, temporal, or spatio-temporal) since they are too local, in the sense that they consider just a few neighbouring video volumes. To overcome these issues, we have developed a multi-scale, hierarchical codebook of BOVs for *densely sampled* videos, which incorporates spatio-temporal *compositions* and their *uncertainties*. This permits the use of statistical inference to recognize the activities. We also note that, in order to measure similarity between a query and a target dataset, it is necessary to use information regarding the *informative* STVs in the video, i.e., the salient foreground objects. To select these space-time regions, we use the information obtained from our hierarchical BOV method, which can be considered as being a context-based spatio-temporal segmentation method.

In this paper we present a hierarchical probabilistic codebook method for action recognition in videos, which is based on STV construction. The method uses both local and global compositional information regarding the volumes, which are obtained by dense sampling at various scales. Similar to other volumetric methods, we do not require background

978-0-7695-4683-4/12 $26.00 © 2012 IEEE
DOI 10.1109/CRV.2012.32

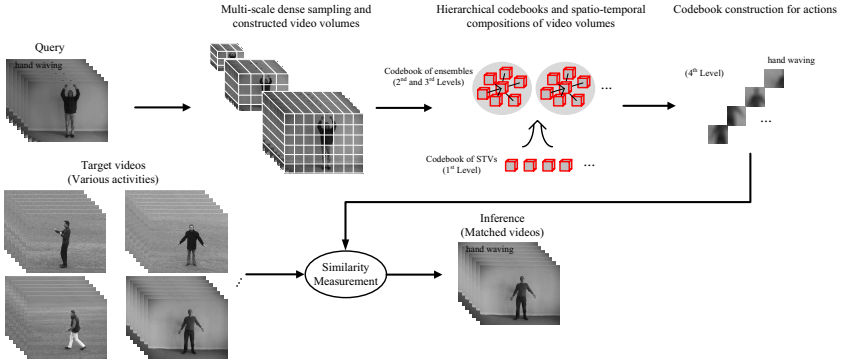


Figure 1: Overview. The goal is to find similar videos in the target set to the query video. The latter is densely sampled at different spatio-temporal scales, after which a four level hierarchical probabilistic codebook is formed using video volumes. Similar video volumes are grouped at the first level of the hierarchy. At the second level, spatio-temporal compositions of the volumes are considered in larger contextual regions to form another codebook. Finally, using temporal correspondences of codewords, the most informative ones are selected for the purpose of classification and used to measure similarity to the target videos.

subtraction, motion estimation, or complex models of body configurations and kinematics. Moreover, the method tolerates variations in appearance, scale, rotation, and movement.

As shown in Figure1, the proposed algorithm consists of two main parts, the hierarchical codebook construction of salient STVs and the inference mechanism for the similarity measurement between salient STVs of the query and target videos. The hierarchical codebook construction has four levels: coding the video to construct STVs and low level probabilistic codebook formation while considering the uncertainties in the STVs; constructing ensembles of video volumes containing a large number of STVs and probabilistic models of their spatio-temporal compositions; high level codebook construction of the ensembles; and finally, analyzing codewords as a function of time in order to construct a codebook of salient regions. The inference mechanism is based on the hierarchical codewords constructed for each query video, and finds the most similar compositions of STVs in the target videos in order match the query and target videos. There are two main differences between our proposed hierarchical approach and previously reported ones. First, the latter are unable to handle uncertainty in the codeword assignments [2], [3]. Second the selection of informative regions is always carried out at the lowest level of the hierarchy.

The main contributions of this paper are as follows:

- We deal with unconstrained videos by incorporating uncertainties during the first level of codeword assignment, which makes the final labelling decision more reliable.
- We introduce a hierarchical codebook structure for action detection and labelling. This is achieved by constructing a probabilistic model of STVs to capture their spatio-temporal configurations.
- We select the salient STVs in the video by analyzing high level codewords that are assigned to each pixel as a function of time. This method is different from conventional background subtraction and salient point detection methods since we use information obtained from a high-level codebook of the video volumes.

In order to evaluate the capability of our approach for action matching and classification we have conducted experiments using three datasets: KTH [4], Weizmann [5] and MSR II [6][1]. Three types of experiments were performed: action matching and retrieval, single dataset video classification, and cross-dataset action recognition. The rest of this paper is organized as follows: section II reviews recent work on action recognition. Section III describes the proposed approach for action recognition and the steps of the algorithm. Section IV discusses the experimental results and finally, section V concludes the paper.

II. Related work

Many studies have focused on the action recognition problem, using different approaches such as human body models, tracking-based methods, and local descriptor methods [1]. Typically, they depend on such image pre-processing as segmentation, object tracking, and background subtraction [7]. Recently, local STVs have been used in the context of BOV models and have shown promising results for action recognition [2], [3], [7]–[13]. In these approaches, video volumes are extracted and quantized in order to form a visual vocabulary. In general, the potential real-time performance of these approaches is related to the number of video volume samples and their associated features [10]. Usually, these features are gradients (spatial, temporal, or spatio-temporal), body landmarks, or color information. The video volumes are constructed either by extracting a limited set of *interest points* or oppositely, by densely sampling the video. In the

[1]http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/

former, due to the sparse nature of the space-time interest points, the method becomes computationally efficient and hence is popular in the action recognition literature [4], [13]–[15]. On the other hand, the selection of appropriate interest points that are *guaranteed* to contain a salient and discriminative motion pattern in their local context is a difficult challenge [16]. In addition, it has been shown recently that densely sampling the video always achieves better results than a sparse set of interest points [17]. Notwithstanding the particular sampling strategy employed, the advantage of using volumetric representations of videos is that it permits the localization and classification of actions using data-driven nonparametric approaches instead of requiring the training of sophisticated parametric models. On the other hand, Boiman et al. [18] have shown that a rather simple nearest-neighbour image classifier in the space of the local image descriptors is equally as efficient as more sophisticated classifiers. Thus the particular classification method chosen is not the major issue, and the main challenge is using the appropriate features for action representation.

In classical BOV approaches, video volumes are grouped based on their similarity, in order to reduce the vocabulary size. Unfortunately, this destroys the compositional information concerning the relationships between volumes [13]. Thus, the likelihood of each video volume is calculated as its similarity to the other volumes in the dataset, without considering the spatio-temporal properties of the neighbouring contextual volumes. This makes the classical BOV approach too dependent on very local data and unable to capture significant spatio-temporal relationships. In addition, it has been shown that detecting actions using an "order-less" BOV will not produce acceptable recognition results [2], [8], [9], [16], [19]. To overcome this challenge, contextual information must be included in the original BOV framework. The solution presented by Boiman and Irani [8] is to densely sample the video and store *all* video volumes, along with their relative locations in *space* and *time*. Thus the likelihood of a query is calculated in a larger space-time contextual region. By also using densely sampled volumes, it is possible to compute the optimal approximation to the likelihood, which yields an accurate label for an action using simple nearest neighbour classifiers [18]. However, the main problem with this approach is that it requires excessive computational time and a considerable amount of memory to store all of the volumes as well as their spatio-temporal relationships.

Several other methods have been proposed to incorporate spatio-temporal structure in the context of BOV. One approach is to use a coarse grid and construct a histogram to subdivide the space-time volumes [14]. Similarly, correlograms were used in [11] to capture the spatio-temporal co-occurrence patterns of the spatio-temporal volumes; however, only the relationship between the two nearest volumes was considered. An alternative is to incorporate contextual information by using a random tree structure [7] to partition the input space and calculate the likelihood of each spatio-temporal region in the video. Otherwise, a hierarchical clustering structure seems to be an attractive way of incorporating the contextual structure of video volumes, as well as preserving their compactness [2], [3]. Thus a modified version of [8] was presented in [3]. It uses a hierarchical approach, in which a two-level clustering method is employed. At the first level, all similar video volumes are clustered. Then clustering is also performed on randomly selected groups of STVs while considering the relationships between the five nearest STVs in space and time. Another hierarchical approach is presented in [2], which attempts to capture the compositional information of a subset of the most discriminative video volumes. In all of these proposed solutions to date, although a higher level of quantization in the action space produces a compact subset of video volumes, it also significantly reduces the discriminative power of the descriptors, an issue addressed in [18].

III. Proposed framework for action recognition

Considering the structure presented in Figure1, our aim is to find the similarity between the query and target videos. Our work is based on the bag of space-time features approach in that a set of STVs is used for measuring similarity. As illustrated in Figure1, the recognition algorithm consists of two main steps: hierarchical codebook construction for densely sampled videos, and an inference mechanism for finding the appropriate action in the target videos. The proposed hierarchical codebook structure has two important characteristics: it codes the compositional information of the video volumes and it selects the most informative regions of the video. Moreover, the uncertainty in the process of grouping similar video volumes is considered in the hierarchical structure, an issue which is always ignored in typical BOV approaches.

A. Multi-scale dense sampling and hierarchical codebook construction

The essence of the method described in this paper is to measure the similarity between STVs in a query video to those in the target set videos. In this section, we first explain the sampling strategy and then describe the construction of the hierarchical codebook. The codebook is intended to reduce the redundancy in the video volumes by retaining the most informative ones, their associated compositional information, and the uncertainties in codeword assignment. The hierarchical codebook structure consists of four layers, which are defined in 3D space-time, and referred to as C^{l_1} to C^{l_4}, as illustrated in Figure2. The first layer contains the STVs obtained from the original video. A codebook is formed at this level to cluster similar densely sampled STVs, called C^{l_1}. At the second level, a large spatio-temporal region is considered around every video volume. We refer

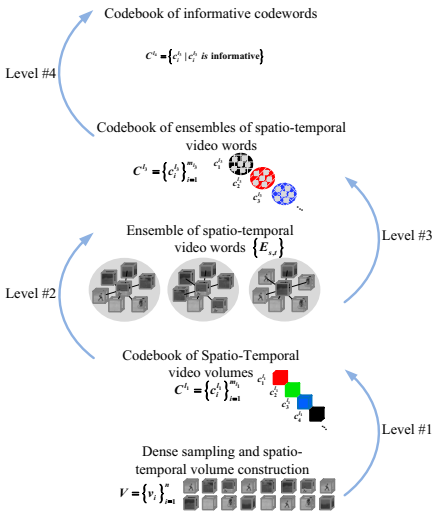


Figure 2: Overview of the four level hierarchical codebook. The first level codebook, C^{l_1}, is constructed to code similar spatio-temporal video volumes obtained from the original video, while considering the uncertainty in codeword assignment. At the second level a larger STV region around each pixel, containing many STVs, is considered to capture the spatio-temporal arrangement of the volumes, called the ensemble of volumes. At the third level similar ensembles are grouped based on the similarity between arrangements of their video volumes and a new codebook is formed, referred to as C^{l_3}. At the fourth level, the codewords obtained from the third level are analyzed as a time series in order to eliminate the non-informative ones.

to this as the ensemble of STVs, $E_{s,t}$, at a point (s,t) in the video. These ensembles are then grouped based on their similarity and an additional codebook, C^{l_3}, is formed. At the fourth level, the codewords obtained from the third level are analyzed as a function of time in order to remove non-informative codewords.

1) Video volume construction: Similar to all BOV approaches, 3D STVs in a video are constructed at the lowest level of the hierarchy. Although there are many methods for sampling the video and for volume construction, we use dense sampling as it has been shown to be superior to the others [8], [17]. Video volumes are constructed assuming a small volume (e.g., $5 \times 5 \times 3$, in which 5×5 is the size of the spatial (image) window and 3 is the depth of the video volume in time) around each pixel in the video. This is performed at several spatial and temporal scales of a Gaussian space-time video pyramid of the original image and produces a large number of STVs for *each* pixel in the video. These are then characterized by a histogram of oriented gradients (HOG) constructed using the quantized gradients of all pixels in each video volume v_i [20]. It should be noted that other more complex descriptors, such as the ones in [12] or the three-dimensional Scale Invariant Feature Transform (SIFT) [21], might further enhance the performance, as the descriptor used here is not invariant with respect to the textures of moving objects. After video volume construction, the STVs are grouped to reduce redundancy and the non-informative volumes are removed.

2) Hierarchical codebook structure: As shown in Figure2, hierarchical codebook construction involves four levels. The main difference between our proposed approach and previously reported hierarchical methods is that the latter are unable to handle uncertainty in the codeword assignments [2], [3]. Moreover, the selection of informative regions is carried out at the highest level of the hierarchy, while it is always performed at the lowest level in classical BOV methods. This resolves the main drawback of such approaches, which is the assumption that the informative regions are spatio-temporally independent, as is the case in interest point selection and pixel-based background subtraction algorithms [22].

Level #1: Codebook of spatio-temporal video volumes. In section III-A1 we described a set of STVs at various spatial and temporal scales using dense sampling. As the number of these volumes is extremely large (for example, about 10^6 in a one minute video) it is advantageous to group similar STVs to reduce the dimensions of the search space. This is commonly performed in all BOV approaches [12], [19]. Here, similar video volumes are also grouped by constructing a codebook. The first observed STV in the video is selected as the first codeword. After that, by measuring the similarity between each observed volume and the codewords already existing in the codebook, either the codewords are updated or a new one is created. Updating is based on the similarity between the newly observed volume and the already existing codewords using weights $w_{i,j}$, see Figure3. Here the Euclidean distance is employed for measuring similarity between volumes and codewords. Thus, the normalized weight of assigning codeword $c_j^{l_1}$ to the video volume v_i is given by:

$$w_{i,j} = \frac{1}{\sum_j \frac{1}{distance\left(v_i, c_j^{l_1}\right)}} \times \frac{1}{distance\left(v_i, c_j^{l_1}\right)} \quad (1)$$

Another parameter used *after* codebook construction is the number of times that a codeword has been observed. During the training period, the codebook is pruned to eliminate those codewords that are either infrequent or very similar to the existing ones. Ultimately, this generates m_{l_1} different codewords that are taken as the labels for the video volumes: $C^{l_1} = \left\{c_1^{l_1}, c_2^{l_1}, ..., c_{m_{l_1}}^{l_1}\right\}$. After the initial codebook formation, each 3D volume, v_i, is assigned to a codeword, $c_j^{l_1}$, with a degree of similarity, $w_{i,j}$ (Figure3). It should be noted that the number of codewords, m_{l_1}, is much less than the number of volumes, n, (for a one minute video, $n = 10^6$, $m_{l_1} = 100$). Moreover, codebook construction can be performed using other clustering methods, such as

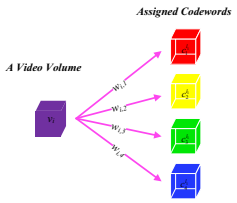


Figure 3: The process of codeword assignment to each spatio-temporal video volume. Each codeword is assigned to a volume with a degree of similarity, $w_{i,j}$.

k-means or mutual information-based clustering [19].

Level #2: Ensemble of spatio-temporal video words. This level in the hierarchy corresponds to the spatio-temporal configuration of video volumes. As mentioned earlier, in order to make the correct decision regarding the query video, it is necessary to analyze the spatio-temporal arrangement of the volumes in a large region [8]. The main drawback of many BOV approaches is that they do not consider the spatio-temporal composition (context) of the volumes. Thus, instead of a single video volume, we consider a large region R around each pixel, containing many video volumes, which capture the spatio-temporal context. Such a set is called an *ensemble* of volumes around the particular pixel in the video. Around each point s in the video at time t, the ensemble of volumes ($E_{s,t}$) is defined as:

$$E_{s,t} = \left\{ v_i^{E_{s,t}} \right\}_{i=1}^{I} \triangleq \{ v_i : v_i \in R_{s,t} \}_{i=1}^{I} \tag{2}$$

where v_j is a spatio-temporal volume, $R_{s,t}$ is a region with pre-defined spatial and temporal radii centered at point (s,t) in the video (e.g., $40 \times 40 \times 30$), and I indicates the total number of volumes in the ensemble. To determine the spatio-temporal compositions of video volumes, we use the *relative* spatio-temporal coordinates of the volume in each ensemble, defined as:

$$x_{v_i^{E_{s,t}}} = \left(C_{v_i} - C_{E_{s,t}} \right) \tag{3}$$

where $x_{v_i^{E_{s,t}}}$ is the relative position of the ith video volume, v_i (in space and time), inside the ensemble of volumes, $E_{s,t}$, for a given point (s,t) in the video. C_{v_i} is the central point of the spatio-temporal volume i in absolute coordinates of 3D space. During codeword assignment at level #1, each volume v_i inside each ensemble $E_{s,t}$ was assigned to a codeword $c^{l_1}_{w_{i,j}^{E_{s,t}}}$ with a weight of $w_{i,j}^{E_{s,t}}$. Therefore, the ensemble is characterized by a set of volume position vectors, codewords and their related weights as follows:

$$E_{s,t} = \left\{ x_{v_i^{E_{s,t}}}, c^{l_1}_{w_{i,1}^{E_{s,t}}}, ..., c^{l_1}_{w_{i,m_{l_1}}^{E_{s,t}}} \right\}_{i=1}^{I} \tag{4}$$

A common approach for calculating similarity between ensembles of volumes is to use the star graph model of [3], [8]. This model uses the joint probability between query and dataset ensembles by decoupling the similarity of the topology of the ensembles from the similarity between the actual video volumes [3]. To avoid such a decomposition, we estimate the *pdf* of the volume compositions in an ensemble. Thus, the probability of a particular arrangement of volumes v inside the ensemble of $E_{s,t}$ is given by:

$$\begin{aligned} &\forall c_i^{l_1} \in \left\{ C^{l_1} \right\} \text{ (first level codebook)} \\ &P_{E_{s,t}}(v) = P\left(x_v, c_1^{l_1}, c_2^{l_1}, ..., c_{m_{l_1}}^{l_1} \right) \\ &\qquad = \sum_{i=1}^{m_{l_1}} P\left(x_v | v = c_i^{l_1} \right) P\left(v = c_i^{l_1} \right) \end{aligned} \tag{5}$$

The first term in the summation in (5), $P\left(x_v | v = c_i^{l_1} \right)$, expresses the topology of the ensembles, and the second term, $P\left(v = c_i^{l_1} \right)$, expresses the similarity of their descriptor values, that is, the weights for the codeword assignment at the first level. We would like to represent each ensemble of volumes by its *pdf*, $P_{E_{s,t}}(v)$. Therefore, given the set of volume positions and their assigned codewords, the probability density function (*pdf*) of each ensemble can be formed using either a parametric model or non-parametric estimation. Here, we approximate the *pdf*s describing each ensemble using histograms.

Level #3: Codebook of ensemble of spatio-temporal video words. At the third level of the hierarchy, similar ensembles of volumes are grouped in order to construct another codebook, that of ensembles of volumes. Using the *pdf* to represent each ensemble of volumes makes it possible to use divergence functions from statistics and information theory as a similarity measure. Here we invoke the Kullback-Leibler (KL) divergence to measure the similarity between two *pdf*s, f and g [23]:

$$KL(f||g) = -\int f(x) \log \left(\frac{g(x)}{f(x)} \right) dx \tag{6}$$

To make the measure symmetric, we take the symmetric KL divergence:

$$d(f||g) = KL(f||g) + KL(g||f) \tag{7}$$

Then the similarity between two ensembles of volumes, E_{s_i,t_i} and E_{s_j,t_j}, is defined as:

$$S\left(E_{s_i,t_i}, E_{s_j,t_j} \right) = e^{-\frac{d^2\left(P_{E_{s_i,t_i}}, P_{E_{s_j,t_j}} \right)}{2\sigma^2}} \tag{8}$$

where $P_{E_{s_i,t_i}}$ and $P_{E_{s_j,t_j}}$ are the *pdf*s of the ensembles E_{s_i,t_i} and E_{s_j,t_j} obtained from the previous level, d is the symmetric KL divergence between the two *pdf*s in (7), and σ is the variance of the KL divergence over all of the ensembles. The third level codebook, $C^{l_3} = \left\{ c_1^{l_3}, c_2^{l_3}, ..., c_{m_{l_3}}^{l_3} \right\}$, is formed using the similarity measurement in (8).

Level #4: Informative codeword selection. Given the codebook C^{l_3} obtained at the third level, each pixel at (s,t) in

the video is assigned to a codeword:

$$p_{s,t} \leftarrow c_i^{l_3} \colon c_i^{l_3} \in C^{l_3} \tag{9}$$

We remove non-informative codewords by analyzing each pixel and its assigned codewords as a function of time. This method was inspired by the pixel-based background model presented in [22], where a time series of quantized color features is created at each pixel from which a compact model of the background is then determined. We adopt the same temporal filtering process as in [22]. The main difference is that we construct these codebooks using different observations for each pixel, which are the assigned codewords obtained from Level #3. Thus, a particular pixel, *s* in a video clip is represented as a sequence of codewords, obtained from (9):

$$P_s = \{p_{s,t} : t \in T\} \tag{10}$$

where *T* is the temporal length of the video.

B. Inference mechanism for a query video (Action matching)

The overall goal of this paper is to find similar videos to a query video in a target set and consequently label them according to the labelled query video. Given the hierarchical codebook for each query video containing a specific activity, the similarity of each pixel in the query video to the target videos is calculated using the ensemble of STVs surrounding it. Finally, the most similar subset of target videos is taken as being similar to the query. Figure4 summarizes the process of determining the hierarchical codebooks and how the inference about the query is obtained.

Hierarchical codebook construction for each query video

For each query video containing a particular action, a_i

- Construct the hierarchical codebook model: $H_{a_i} = \{C_{a_i}^{l_1}, C_{a_i}^{l_2}, C_{a_i}^{l_3}, C_{a_i}^{l_4}\}$

Matching target videos

For each target video, g,

- Densely sample the video at all scales and construct spatio-temporal volumes

For each subset of query videos, containing a particular action a_i

- Assign each video volume in the target to $C_{a_i}^{l_1}$
- Construct an ensemble of volumes at each particular pixel, $E_{s,t}^g$
- Measure similarity of the ensembles to $C_{a_i}^{l_4}$ using (8): $s_{E_{s,t}^g, a_i} = \underset{j}{Max}\, S\left(E_{s,t}^g, c_j^{l_4} : c_j^{l_4} \in C_{a_i}^{l_4}\right)$
- Calculate the likelihood of the target: $s_{g,a_i} = \sum_{s,t} s_{E_{s,t}^g, a_i}$

Action determination:

- The ensemble $E_{s,t}^g$ contains action a_{i^*}: $s_{E_{s,t}^g, a_{i^*}} \geq \gamma, \qquad i^* = \arg \underset{i}{Max}\left(s_{E_{s,t}^g, a_i}\right)$

Target classification:

- The target contains action a_{i^*}: $i^* = \arg \underset{i}{Max}\ (s_{g,a_i})$

Figure 4: The complete algorithm for similarity measurement between query and target videos.

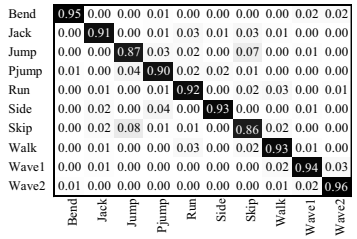
IV. Experimental results

In order to measure the capabilities of the proposed method for action recognition, the algorithm was tested on three different datasets: KTH [4], Weizmann [5] and MSR II [6]. The Weizmann and KTH datasets are the standard benchmarks in the literature used for action recognition. The Weizmann dataset consists of ten different actions performed by nine actors, and the KTH action data set contains six different actions, performed by twenty-five different persons in four different scenarios (indoor, outdoor, outdoor at different scales, outdoor with different clothes). The MSR II consists of 54 video sequences, recorded in different environments with cluttered backgrounds in crowded scenes, and contains three types of actions similar to the KTH: boxing, hand clapping, and hand waving. We evaluate our approach for three different scenarios: action matching and retrieval, single dataset video classification, and cross-dataset action detection. Here, single dataset classification implies that both target and query videos are selected from the same dataset, while cross-dataset recognition assumes that the query and target videos are selected from different datasets.

Video matching and classification are performed using KTH and Weizmann datasets, which are single-person, single-activity videos. Although they were collected in controlled environments, we use them to compare with the current state-of-the-art. For cross-dataset action recognition, we use the KTH dataset as the query set, while the target videos are selected from the more challenging MSR II dataset. Our experiments demonstrate the effectiveness of our hierarchical codebook method for action recognition in various categories.

A. Action matching and retrieval

Since our proposed method is a video-to-video matching framework, it is not necessary to have a training sequence. This means that we can select one labelled query video for each action, and find the most similar one to it to perform the labelling. For the Weizmann dataset, we used one person for each action as the query video and the rest (eight other persons) as the target sets. This was done for all persons in the dataset and the results were averaged. The confusion matrix for the Weizmann dataset is shown in Figure5, achieving an average recognition rate of 91.7% over all 10 actions. The columns of the confusion matrix represent the instances to be classified, while each row indicates the corresponding classification results. We carried out the same experiment on the KTH dataset. The confusion matrix is shown in Figure5. The average recognition rate was 84.33% over all 6 actions. The results indicate that the method proposed in this paper outperforms state-of-the-art approaches, even though the former requires no background/foreground segmentation and tracking. The average accuracy of the other methods is presented in TableI. The overall results on the Weizmann dataset are better than those on the KTH dataset. This is predictable, since the Weizmann dataset contains videos with more static backgrounds and more stable and discriminative actions than the KTH dataset.



(a) Weizmann dataset (b) KTH dataset

Figure 5: Confusion matrices for single video action matching, a) Weizmann dataset, b) KTH dataset. A single video is used as the query to which the other videos in the dataset are matched.

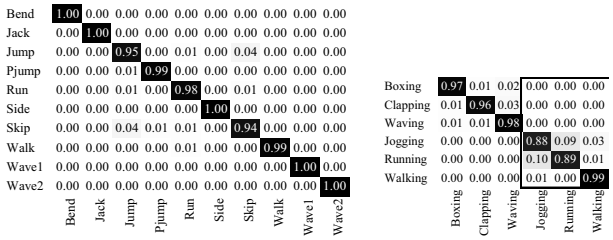
Table I: Action recognition comparison with the state-of-the-art using single video action matching

Method	Dataset	
	KTH	Weizmann
Our method	84.33	91.7
Thi et.al. [24]	77.17	88.6
Seo et.al. [12]	69	78

B. Single Dataset Action classification

In order to make an additional quantitative comparison of our algorithm with the state-of-the-art, we have extended it to the action classification problem. This refers to the more classical situation in which we use a *set of query videos* instead of just a single one, as discussed previously. We have evaluated our algorithm's ability to apply the correct label to a given video sequence, when both the training[2] and target datasets are obtained from the same dataset. We tested the Weizmann and KTH datasets, and applied the standard experimental procedures in the literature. For the Weizmann dataset, the common approach for classification is to use leave-one-out cross-validation, i.e., eight persons are used for training and the videos of the remaining person are matched to one of the ten possible action labels. Consistent with other methods in the literature, we mixed up the four scenarios for each action in the KTH dataset and followed the standard experimental procedure for this dataset [6], in which 16 persons are used for training and nine for testing, done randomly 100 times. Then, we calculated the average performance over these random splits. The confusion matrix for the Weizmann dataset is reported in Figure6 and the average recognition rate is 97% over all 10 actions in the leave-one-out setting. As expected from earlier experiments reported in the literature, our results indicate that the "skip" and "jump" actions are easily confused, as they appear visually similar. On the KTH dataset, we achieved an average recognition rate of 94.5% for the six actions, shown in the confusion matrix in Figure6. As observed from Figure6, the primary confusion occurs between jogging and running,

[2]Although our method does not require any specific training sequences, we refer to the query video dataset as the training set for consistency with the literature.



(a) Weizmann dataset (b) KTH dataset

Figure 6: Confusion matrices for the action classification, a) Weizmann dataset, b) KTH dataset.

Table II: Comparison of action recognition with the state-of-the-art

Method	Dataset	
	KTH	Weizmann
Our method	94.5	98.5
Bregonzio et al. [9]	93.17	96.6
Yao et al. [7]	93.5	97.8
Thi et.al. [24]	94.67	98.9
Seo et.al. [12]	95.1	97.5
Wang et al. [14]	93.8	-
Liu et al. [19]	94.2	-
Tian et al. [25]	94.5	-

which is also problematical for the other approaches. Obviously, this is due to the inherent similarity between the two actions. The recognition rate was also compared to other approaches (see TableII). Comparing our results with those of the state-of-the-art, we observe that they are similar, even though we do not require any background/foreground segmentation and tracking.

C. Cross-dataset action matching and retrieval

Similar to other approaches for action recognition [25], we use cross-dataset recognition to measure the robustness and generalization capabilities of our algorithm. In this paradigm, the query videos are selected from a specific dataset (in our experiments the KTH dataset) and the targets from another(MSR II), so that we compare similar actions performed by different persons in different environments. We selected three classes of actions from the KTH dataset as the query videos: boxing, hand waving, and hand clapping, including 25 persons performing each action. A hierarchical codebook was formed for each action category and the query was matched to the target videos. We varied the detection threshold to obtain the precision/recall curves for each action type, as shown in Figure7. This achieved an overall recognition rate of 79.6%, which is better than the state-of-the-art (see TableIII).

Table III: Comparison of action recognition with the state-of-the-art

Method	Accuracy (%)
Our method	79.6
Tian et al. [25]	78.8
Yuan et al. [15]	59.6

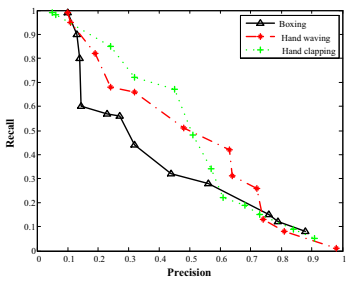


Figure 7: The precision-recall curves for cross-dataset action recognition using the hierarchical codebook structure.

V. Conclusion and Future Work

We have presented a new hierarchical approach based on spatio-temporal volumes for the challenging problem of human action recognition in videos. Our approach is an extension to the conventional BOV approaches since we construct a hierarchical representation of informative video volumes and their compositional relationships. The hierarchical structure consists of four levels:

1) Coding a video using spatio-temporal volumes to produce a low-level codebook.
2) Constructing an ensemble of video volumes and representing their structure using probabilistic modeling of the relative compositions of the spatio-temporal volumes.
3) High level codebook construction of ensemble volumes.
4) Analysis of the codewords as a function of time in order to construct a codebook of salient regions.

Given a single query video (an example of a particular activity), the method computes the similarity of each pixel in each frame of the target videos to the query, and finds the subset of target videos which are similar to that query. This is accomplished by analyzing a relatively large contextual region around the pixel, while considering the compositional structure using a probabilistic framework. The algorithm was tested on three popular benchmarks, KTH, Weizmann, and MSR. We showed that the algorithm is effective and robust, in both action-matching and cross-dataset recognition tasks. Moreover, the results are highly competitive with state-of-the-art methods. However, a major advantage of our approach is that it does not require any feature analysis, background/foreground segmentation and tracking, and is susceptible to on-line real-time analysis. The proposed video method can easily be extended to multi-action retrieval and action localization by modifying the inference mechanism.

Acknowledgments

The authors would like to acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the McGill International Doctoral Awards (MIDA).

References

[1] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.*, vol. 28, no. 6, pp. 976–990, 2010.
[2] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *CVPR*, 2010, pp. 2046–2053.
[3] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal localization and categorization of human actions in unsegmented image sequences," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1126–1140, 2011.
[4] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *ICPR*, vol. 3, 2004, pp. 32–36.
[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.
[6] J. Yuan, Z. Liu, and Y. Wu, "Discriminative video pattern search for efficient action detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1728–1743, 2011.
[7] A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *CVPR*, 2010, pp. 2061–2068.
[8] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vision*, vol. 74, no. 1, pp. 17–31, 2007.
[9] M. Bregonzio, G. Shaogang, and X. Tao, "Recognising action as clouds of space-time interest points," in *CVPR*, 2009, pp. 1948–1955.
[10] Y. Ke, R. Sukthankar, and M. Hebert, "Volumetric features for video event detection," *Int. J. Comput. Vision*, vol. 88, no. 3, pp. 339–362, 2010.
[11] S. Savarese, A. DelPozo, J. C. Niebles, and F.-F. Li, "Spatial-temporal correlatons for unsupervised action classification," in *WMVC*, 2008, pp. 1–8.
[12] H. Seo and P. Milanfar, "Action recognition from one example," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 867–882, 2011.
[13] J. C. Niebles, H. C. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, 2008.
[14] J. Wang, Z. Chen, and Y. Wu, "Action recognition with multiscale spatio-temporal contexts," in *CVPR*, 2011, pp. 3185–3192.
[15] J. Yuan, Z. Liu, and Y. Wu, "Discriminative subvolume search for efficient action detection," in *CVPR*, 2009, pp. 2442–2449.
[16] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *CVPR*, 2009, pp. 1446–1453.
[17] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.
[18] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," *CVPR*, pp. 1992–1999, 2008.
[19] J. Liu and M. Shah, "Learning human actions via information maximization," in *CVPR*, 2008, pp. 1–8.
[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1, 2005, pp. 886–893.
[21] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *International conference on Multimedia*, 2007, pp. 357–360.
[22] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 172–185, 2005.
[23] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York: Springer, 2006.
[24] T. H. Thi, L. Cheng, J. Zhang, L. Wang, and S. Satoh, "Integrating local action elements for action analysis," *Compt. Vis. Image Und.*, vol. 116, no. 3, pp. 378–395, 2012.
[25] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *IEEE Trans. Syst., Man, Cybern. C*, vol. PP, no. 99, pp. 1–11, 2011.