Video Classification using Semantic Concept Co-occurrences

Shayan Modiri Assari Amir Roshan Zamir Mubarak Shah
Center for Research in Computer Vision, UCF

Abstract

We address the problem of classifying complex videos based on their content. A typical approach to this problem is performing the classification using semantic attributes, commonly termed concepts, which occur in the video. In this paper, we propose a contextual approach to video classification based on Generalized Maximum Clique Problem (GMCP) which uses the co-occurrence of concepts as the context model. To be more specific, we propose to represent a class based on the co-occurrence of its concepts and classify a video based on matching its semantic co-occurrence pattern to each class representation. We perform the matching using GMCP which finds the strongest clique of co-occurring concepts in a video. We argue that, in principal, the co-occurrence of concepts yields a richer representation of a video compared to most of the current approaches. Additionally, we propose a novel optimal solution to GMCP based on Mixed Binary Integer Programming (MBIP). The evaluations show our approach, which opens new opportunities for further research in this direction, outperforms several well established video classification methods.

1. Introduction

Classification of complex videos is an active area of research in computer vision. Despite the complicated nature of unconstrained videos, they can be described as a collection of simpler lower-level concepts, such as *candle blowing*, *walking*, *clapping*, etc. Therefore, a typical approach to video categorization is to first apply concept detectors to different segments of the test video and form a histogram of concepts occurring therein. Next, a trained classifier determines which class the histogram may belong to.

In this paper, we propose an approach to complex video classification that models the context using the pairwise co-occurrence of concepts. Generalized Maximum Clique Problem is useful in situations where there are multiple potential solutions for a number of subproblems, along with a global criterion to satisfy. We use GMCP in order to select a set of concepts in different clips of the video in a way that they are holistically in agreement. Thus, a concept that is out of context in the whole video does not appear in our results, while they are common when the concept detection is done in an individual manner. Also, we propose a new solution to GMCP using Mixed Binary Integer Programming.

Figure 1. We represent a video category based on the co-occurrences of the semantic concepts happening therein and develop a classifier based on cliques of concepts. The nodes represent semantic concepts and the edges denote the strength of co-occurrence factor between them for a sample video class.

We develop a class specific co-occurrence model and propose a method which uses the GMCP as the classifier and the class-specific co-occurrence models learnt from a training set as the representation of the classes (shown in Fig. 1). We argue that this representation is essentially more *semantically meaningful* and *fast* in computation compared to the traditional representations, such as the collection of concept histograms of class videos [9]. We show that the proposed classification method significantly outperforms the baseline in particular for videos which include enough contextual cues.

Several methods for concept detection and classification have been developed during the past few years. Liu and Huet [12] developed a method for automatic refinement of concept detectors using the massive amount of available data on the web. Wei et al. [15] proposed a concept-driven approach to fusion of multi-modal information for an efficient video search. Izadinia and Shah [8] present a method for modeling the relationship between low-level events (concepts) in a framework based on latent SVM. Wang et al. [5] developed a fusion based method for building an effective training set for video categorization. Jiang

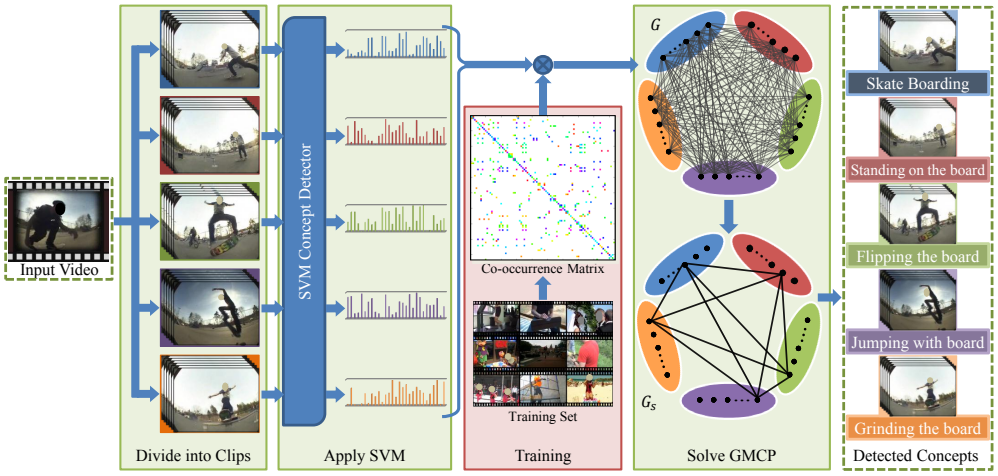


Figure 2. The block diagram of our concept detection method. The testing and training steps are shown in green and red respectively.

et al. [13] proposed a contextual approach to concept detection using conditional random fields. Each node in the defined graph represents a concept within one clip, and the detection probabilities are improved leveraging boosting for fusion. Discriminative Fusion Model (DMF) [7] is another two-layer approach to contextual concept detection; first individual detectors are applied to a clip, then the detection scores of all individual detectors are fed into a SVM to determine the concept label based on all scores.

What differentiates our approach from the aforementioned methods is that we represent a class directly based on co-occurrences of semantic concepts happening therein; this is different from the majority of the existing methods which focus on the *occurrence* of semantic cues more than their *co-occurrence*. Also, unlike the existing methods such as [11], we classify a video directly based on discovering the underlying co-occurrence pattern therein and fitting it to the learnt co-occurrence patterns of different classes. In this context, we migrate from the conventional vector-representations to the richer matrix-representation which is fundamental to the rest of our clique-based framework. Moreover, many of the aforementioned methods, such as [13, 7], perform the fusion of concepts within one shot of the video. Our method not only incorporates the relationship of all concepts in one clip, but also it fuses the information among different clips of the video. This is in particular important for contextual concept detection in long videos.

The contributions of this paper can be summarized as: 1. A new representation for video categories based on the co-occurrence of their semantic concepts 2. A novel complex video classification method based on the proposed representation and GMCP 3. A novel optimal GMCP solver using Mixed Binary Integer Programming (MBIP). [1]

[1]More details available at the project website: http://crcv.ucf.edu/projects/GMCP_Classifier/.

2. Contextual Concept Detection using GMCP

The block diagram of the proposed concept detection method is shown in fig. 2. In training, the probability of concept co-occurrences are computed from an annotated training set and saved in a reference co-occurrence matrix.

In testing, the query video is divided into clips of fixed size. Let k and h denote the number of defined concepts and number of clips in the test video respectively. We apply k trained concept detectors to each clip and use the resulting $k \times h$ confidence values along with a reference co-occurrence matrix to form the graph G. Each clip is represented by a cluster of k nodes representing concepts in that clip in G, and the edge weights between nodes specify the probability of co-occurrence of the corresponding concepts in the test video based on both SVM confidences and training data. By solving GMCP (which selects one node from each cluster) for the graph G, the set of concepts which is in maximal contextual agreement is found.

2.1. Context Model: Co-occurrence of Concepts

We use the pairwise co-occurrence of concepts as our means of capturing the context in a video. We define the $k \times k$ reference co-occurrence matrix based on the conditional probability of coincidence of concepts:

$$\Phi(a,b) = p(a|b) = \frac{\#(a,b)}{\#(b)}, \tag{1}$$

where $\#(a)$ is the number of training videos which include the concept a, and $\#(a,b)$ represents the number of videos in which both concepts a and b occur. For the self co-occurrence elements, i.e. $\Phi(a,a)$, the numerator term $\#(a,a)$ is equivalent to the number of videos in which concept a occurs more than once.

The element $\Phi(a,b)$ is actually equivalent to the conditional probability $p(a|b)$ which is the probability of concept

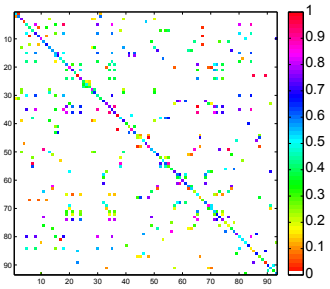


Figure 3. Extracted reference co-occurrence matrix, Φ, from the annotations of TRECVID-MED 2011-2012 videos. We have total number of $k = 93$ concepts such as: *person walking, clapping, animal eating, vehicle moving*.

a happening in a video given concept b happens. Defining the co-occurrence matrix using the conditional probability has a number of advantages: it does not penalize the *co-occurrence* of concepts which tend to *occur* less often. Also, the resulting co-occurrence matrix is asymmetric; this is of particular importance as in principle, the chance of concept a happening in a video given concept b happens is not necessarily the same as vice versa. Fig. 3 shows the computed co-occurrence matrix from our training set which includes videos with manually annotated concepts.

We used *pairs* of concepts for modeling the context as adopting a higher order, such as triplets or quadruples of concepts, would require a substantially larger training set[2]. In the next subsection, we explain how GMCP can effectively leverage the pairwise reference co-occurrence matrix to identify the incorrectly detected concepts in a test video and assign them the right labels.

2.2. Improving Concept Detection using GMCP

We define the input to our contextual concept detection method as the graph $G = \{\mathbf{V}, \mathbf{L}, \boldsymbol{w}\}$ where $\mathbf{V}$, $\mathbf{L}$, and $\boldsymbol{w}$ represent the set of nodes, edges and edge weights respectively. The nodes in $\mathbf{V}$ are divided into disjoint clusters where each cluster, C, represents one clip of the test video, and the nodes therein denote the potential concepts for that particular clip. Thus, $C_j = \{\alpha_1^j, \alpha_2^j, \alpha_3^j, \ldots, \alpha_k^j\}$ where α_i^j represents the i^{th} concept of j^{th} clip. $\mathbf{L}$ includes the edges between all of the possible pairs of nodes in $\mathbf{V}$ as long as they do not belong to the same cluster. We define $\boldsymbol{w}$ as:

$$\boldsymbol{w}(\alpha_i^j, \alpha_l^m) = \overbrace{\Phi(\alpha_i^j, \alpha_l^m)}^{context}.\overbrace{\psi(\alpha_l^m)}^{content}, \qquad (2)$$

where $\Phi(\alpha_i^j, \alpha_l^m)$ determines the contextual agreement between the two concepts α_i^j and α_l^m which are from two different clips. We apply k trained concept detector to each clip; $\psi(\alpha_l^m)$ denotes the confidence value of the l^{th} concept detector applied on the m^{th} clip. Note that the defined edge weight $\boldsymbol{w}(\alpha_i^j, \alpha_l^m)$ is equivalent to the probability of α_i^j and α_l^m occurring in the test video: $p(\alpha_i^j \cap \alpha_l^m) = p(\alpha_i^j | \alpha_l^m).p(\alpha_l^m)$, where the first term is from the reference co-occurrence matrix and the second is the SVM confidence value (after proper calibration and normalization). Therefore, a larger edge weight implies a higher probability for its parent concepts, α_i^j and α_l^m, to occur in the test video.

[2]Assume the training set includes λ videos with average number of μ clips in each. $\binom{\mu}{o}.\lambda$ relationships for concept sets of order o can be extracted from our training set, which are used for modeling k^o possible concept sets. The ratio of $\binom{\mu}{o}.\lambda/k^o$ sharply drops as o increases, therefore, a substantially large training set is required when o is increased.

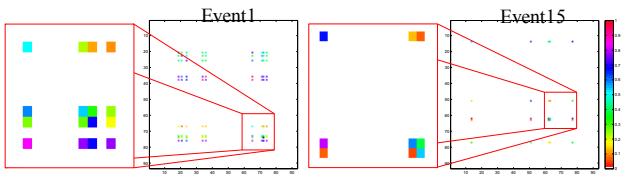


Figure 4. The co-occurrence matrices extracted for two sample classes of TRECVID-MED 2011-2012 datasets. Part of each matrix is magnified on the left to show more details. The class specific co-occurrence matrices are sparser compared to the one shown in fig. 3 as some of the concepts occur in particular classes only.

In order to perform the concept detection, one concept needs to be assigned to each clip of the test video. Therefore, a feasible solution to this problem can be represented by a subgraph of G. We call this subgraph $G_s = \{\mathbf{V}_s, \mathbf{L}_s, \boldsymbol{w}_s\}$ where $\mathbf{V}_s$ must include one and only one node from each clip. Hence, $\mathbf{V}_s \subset \mathbf{V}$, and $\mathbf{L}_s$ and $\boldsymbol{w}_s$ are the subsets of $\mathbf{L}$ and $\boldsymbol{w}$ which the nodes in $\mathbf{V}_s$ induce. A sample graph G and one of its feasible subgraphs, G_s, are shown in Fig. 2.

We define the following utility function which assigns a score to the feasible solution G_s:

$$U(G_s) = \frac{1}{h.(h-1)} \sum_{p=1}^{h} \sum_{q=1, q\neq p}^{h} \boldsymbol{w}(\mathbf{V}_s(p), \mathbf{V}_s(q)), \qquad (3)$$

which is the sum of the edge weights of the subgraph G_s.

All of the possible pairwise relationships between different concepts in different clips are incorporated in eq. 3; therefore, by solving the following optimization problem, the set of contextually consistent concepts is found:

$$G_s^* = \arg\max_{G_s} U(G_s) = \arg\max_{\mathbf{V}_s} \sum_{p=1}^{h} \sum_{q=1, q\neq p=1}^{h} \boldsymbol{w}(\mathbf{V}_s(p), \mathbf{V}_s(q)). \qquad (4)$$

We use GMCP for solving the above combinatorial optimization problem.

Generalized Maximum Clique Problem. The objective of Generalized Maximum Clique Problem is finding a subgraph within a complete graph with clustered nodes in a way that the sum of its edge weights is optimized [6]. More formally, the input graph to GMCP is defined

as $G = \{\mathbf{V}, \mathbf{L}, \boldsymbol{w}\}$ where the nodes in $\mathbf{V}$ are divided into disjoint clusters and no intra cluster edge exists. GMCP finds the subgraph $G_s = \{\mathbf{V}_s, \mathbf{L}_s, \boldsymbol{w}_s\}$ within G such that it selects exactly one node from each cluster to be included in $\mathbf{V}_s$ and the summation of values in $\boldsymbol{w}_s$ is maximized. As can be inferred from the definition of GMCP, finding the generalized maximum clique in our concept detection input graph G will essentially solve the optimization problem of eq. 4. Therefore, we solve GMCP to our input graph G in order to discover the set of concepts in the test video.

Several suboptimal solutions for GMCP have been developed in different fields such as Communications, Biology and Data Association [3, 10, 18, 6]. In sec. 4, we propose an optimal solution to GMCP.

Note that in this section, we detect concepts in a way that they are contextually consistent, regardless of what class the video belongs to. In the next section, we propose to make the co-occurrence matrix class-specific and utilize it for video classification using GMCP.

3. Video Classification using GMCP

We propose a representation for a video class based on the co-occurrence of its concepts and develop a method which uses GMCP as the classifier. We define the class specific co-occurrence matrix Φ' as:

$$\Phi'(a, b, \epsilon) = p(a|b, \epsilon) = \frac{\#_\epsilon(a, b)}{\#_\epsilon(a)}, \tag{5}$$

where $\#_\epsilon(a)$ is the number of training videos of class ϵ which include concept a, and $\#_\epsilon(a, b)$ represents the number of training videos of class ϵ in which both concepts a and b occur. Therefore, $\Phi'(., ., \epsilon)$ contains the pattern of concept co-occurrences for class ϵ. Fig. 4 shows the co-occurrence matrices trained for two classes of TRECVID11,12-MED dataset.

Representing an complex video class using the co-occurrence of its concepts has several advantages over the traditional representations such as histogram of concepts:

1. *Speed*: finding the representation is almost instantaneous as it requires counting the coincidence of k concepts in the training set.

2. *Concept Interactions*: It is based on discovering the correlation of concepts, yet it captures what concepts typically occur in a class. This is different from most of the existing methods such as histogram of concepts which are mainly based on *occurrence* information.

3. *Semantics*: it is semantically meaningful. This enables using alternative resources, such as web documents or YouTube video labels, to be used for computing the representation when video annotations are not available.

In order to preform the classification using GMCP, we define the input graph $G' = \{\mathbf{V}, \mathbf{L}, \boldsymbol{w}', \epsilon\}$ to represent the test video. The set of nodes, $\mathbf{V}$, and edges, $\mathbf{L}$, are the same as in G, by definition. We define the edge weights as:

$$\boldsymbol{w}'(\alpha_i^j, \alpha_l^m, \epsilon) = \Phi'(\alpha_i^j, \alpha_l^m, \epsilon).\psi(\alpha_l^m), \tag{6}$$

where ϵ is the class its co-occurrence matrix is being used for computing the edge weights. Hence, assuming E classes exist in our dataset, we form E different input graphs G' for a test video. Similar to the concept detection method described in sec. 2.2, a feasible solution to the classification problem can be represented by a subgraph of G' which we define as $G'_s = \{\mathbf{V}_s, \mathbf{L}_s, \boldsymbol{w}'_s, \epsilon\}$. The class-assignment utility function of a feasible solution is defined as:

$$U'(G'_s, \epsilon) = \frac{1}{h.(h-1)} \sum_{p=1}^{h} \sum_{q=1, p\neq q}^{h} \boldsymbol{w}'(\mathbf{V}_s(p), \mathbf{V}_s(q), \epsilon), \tag{7}$$

which assigns E different scores to the feasible solution G'_s. Each score represents how well the feasible solution fits the co-occurrence pattern of the corresponding class as well as the confidence score of the concept detectors. Thus, by solving the following optimization problem, we can find which class the test video belongs to as well as its concepts:

$$\{G^*_s, \epsilon^*\} = \arg\max_{G'_s, \epsilon} U'(G'_s, \epsilon), \tag{8}$$

where ϵ^* and G^*_s represent the found class and the optimal subgraph found using the co-occurrence matrix of class ϵ^*.

In summary, we represent a test video E times using E different co-occurrence matrices and solve GMCP for each. The class which yields the highest score is selected as the recognized class. In sec. 5, we show that this approach outperforms the existing methods such as using a multiclass SVM classifier.

4. Solving GMCP using Mixed Binary Integer Programming (MBIP)

A number of approximate solutions have been proposed for GMCP, such as local neighborhood search and branch-and-cut [6, 18, 3, 17]. However, no efficient method for solving GMCP in an optimal manner has been developed to date. We propose an MBIP solution to GMCP which guarantees the optimal answer. Finding the optimal solution is in particular important for us as in the GMCP-based classifier, the class-specific co-occurrence matrices are typically sparse; this makes getting stuck in suboptimal regions more likely as there is no gradient in the solution space to preform the descending on.

First, we formulate GMCP through Binary Integer Programming; then, we show the problem can be reduced to Mixed Binary Integer Programming:

4.1. Solving GMCP using Binary Integer Programming (BIP)

The standard form of a Binary Integer Programming problem is [1]:

$$\begin{cases} \text{maximize} & \mathbf{W}^T\mathbf{X}, \\ \text{subject to} & \mathbf{AX} = \mathbf{B}, \\ \text{and} & \mathbf{MX} \leq \mathbf{N}, \end{cases}$$

where $\mathbf{X}$ is a column vector of a number of boolean variables which is supposed to represent a feasible solution of GMCP. Therefore, for each node and edge in our GMCP input graph G, we put a variable in $\mathbf{X}$ which take the value of 1 if the corresponding node or edge is included in the feasible solution G_s. Let ν_i^j denote the boolean variable of i^{th} node in j^{th} cluster, and ε_{mn}^{ji}be the boolean variable representing the edge between the nodes α_i^j and α_n^m. Note that the defined edge weights in eq. 2 are asymmetric, i.e. $\boldsymbol{w}(\alpha_i^j, \alpha_n^m) \neq \boldsymbol{w}(\alpha_n^m, \alpha_i^j)$. However, if the nodes α_n^m and α_i^j are selected to be in G_s, then both $\boldsymbol{w}(\alpha_i^j, \alpha_n^m)$ and $\boldsymbol{w}(\alpha_n^m, \alpha_i^j)$ are included in G_s. Hence, both edges can be represented by a single edge with the weight $\boldsymbol{w}(\alpha_i^j, \alpha_n^m) + \boldsymbol{w}(\alpha_n^m, \alpha_i^j)$. Therefore, we put one variable for such pairs in $\mathbf{X}$, so ε_{mn}^{ji}=ε_{ji}^{mn}, and we use them interchangeably. $\mathbf{X}$ which is of the size $(h.k + \binom{h}{2}.k^2) \times 1$ is defined to have the following general form:

$$\mathbf{X} = [\nu_1^1, \nu_2^1, ..., \nu_{k-1}^h, \nu_k^h, \varepsilon_{11}^{21}, \varepsilon_{11}^{22}, \varepsilon_{11}^{23}, ..., \varepsilon_{(h-1)k}^{h(k-1)}, \varepsilon_{(h-1)k}^{hk}]^T.$$

$\mathbf{X}$ should satisfy the following three constrains in order to be a valid GMCP solution:

Constraint 1 enforces that the summation of node variables of one cluster has to be one, which ensures that one and only one node from each cluster is selected:

$$\{\forall j | 1 \leq j \leq h\} : \sum_{i=1}^{k} \nu_i^j = 1. \tag{9}$$

Constraint 2 states if one node is selected to be in G_s, then exactly $(h-1)$ of its edges should be included in G_s (this is because each node should be connected to all other $(h-1)$ clusters):

$$\{\forall m, n | 1 \leq m \leq h, 1 \leq n \leq k\} : \sum_{i=1}^{h} \sum_{j=1}^{k} \varepsilon_{mn}^{ij} = \nu_n^m.(h-1). \tag{10}$$

Constraint 3 ensures if an edge is included in G_s, then the variables of the nodes incident to it are 1 and vice versa:

$$\{\forall m, n, i, j | 1 \leq m, j \leq h, 1 \leq n, i \leq k\} : \nu_n^m \wedge \nu_i^j = \varepsilon_{mn}^{ji}. \tag{11}$$

Any $\mathbf{X}$ which satisfies the aforementioned three constrains represents a valid solution to GMCP. However, the second constraint implies the third, since it does not allow selecting an edge without selecting its respective nodes. Therefore, the first two constraints are sufficient.

We enforce these constrains using $\mathbf{AX} = \mathbf{B}$ term in BIP formulation. $\mathbf{A}$ and $\mathbf{B}$ are matrices of sizes $(h + h.k) \times (h.k + \binom{n}{2}.k^2)$ and $(h + h.k) \times 1$ respectively. Let $\mathbf{B}$ be:

$$\{1 \leq i \leq h + h.k\} : \mathbf{B}(i) = \begin{cases} 1 & \text{if } 1 \leq i \leq h, \\ 0 & \text{otherwise.} \end{cases}$$

We enforce constrains 1 in $\mathbf{A}$ using the following equation; the first h row of $\mathbf{A}$ enforce the summation of the node variables of the h clusters to be one:

$$\{1 \leq i \leq h, 1 \leq j \leq (h.k + \binom{n}{2}.k^2)\} :$$
$$\mathbf{A}(i,j) = \begin{cases} 1 & \text{if } \lfloor \frac{j-1}{k} \rfloor = i - 1, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $\lfloor . \rfloor$ indicates the floor function. The following equation enforces constraint 2 using the rest of the rows of $\mathbf{A}$:

$$\{(h+1) \leq i \leq (h + h.k), 1 \leq j \leq (h.k + \binom{n}{2}.k^2)\} :$$
$$\mathbf{A}(i,j) = \begin{cases} -(h-1) & \text{if } j = i - h, \\ 1 & \text{if } \mathbf{A}(i,j) \in \mathcal{X}, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

The i^{th} row of $\mathbf{A}$ enforces constraint 2 for the $(i-h)^{th}$ element in $\mathbf{X}$. In the matrix multiplication $\mathbf{AX} = \mathbf{B}$, the element $\mathbf{A}(i, i-h)$ is multiplied with the node variable $\nu_{(i-h-1 \mod k)+1}^{\lfloor \frac{i-h-1}{k} \rfloor + 1}$. Hence, we set $\mathbf{A}(i, i-h) = -(h-1)$, and define $\mathcal{X}$ as the set of elements corresponding to the edges of the node $\nu_{(i-h-1 \mod k)+1}^{\lfloor \frac{i-h-1}{k} \rfloor + 1}$.

$\mathbf{W}$ is a matrix with the same size as $\mathbf{X}$ defined as:

$$\{1 \leq i \leq (h.k + \binom{n}{2}.k^2)\} :$$
$$\mathbf{W}(i) = \begin{cases} 0 & \text{if } 1 \leq i \leq h.k, \\ w(\nu_j^l, \nu_n^m) + w(\nu_n^m, \nu_j^l) & \text{otherwise,} \end{cases} \tag{14}$$

where in the lower equation, we assumed i^{th} element in $\mathbf{X}$ and $\mathbf{W}$ corresponds to the edge variable ε_{mn}^{lj}; hence we put $\mathbf{W}(i)$ equal to the sum of the corresponding edge weights.

Finally, by maximizing $\mathbf{W}^T\mathbf{X}$, the optimal vector $\mathbf{X}$ which satisfies the above constrains is found. Therefore, we have formulated GMCP as a BIP problem which is guaranteed to yield the optimal GMCP answer as a BIP problem can be optimally solved.

4.2. Reduction from BIP to Mixed Binary Integer Programming (MBIP)

All the variables are forced to be binary in a BIP problem. We show that forcing the node variables to be binary guarantees the convergence of edge variables to binary values provided they range from 0 to 1; therefore, our problem can be reduced to Mixed Binary Integer Programming (MBIP) which is less complex than BIP:

Proposition 1
if $\{\forall m,n,i,j | 1 \leq m,j \leq h, 1 \leq n,i \leq k\} : \nu_n^m \in \{0,1\}$ *and* $0 \leq \varepsilon_{mn}^{ji} \leq 1$, *then* $\varepsilon_{mn}^{ji} \in \{0,1\}$.

Constraint 1 is still valid, which enforces exactly one node from each cluster should be selected. Constraint 2 implies a non-zero edge cannot be connected to a zero node; therefore, combined with constraint 1, it yields all non-zero edges in one cluster should belong to one node. Additionally, no more than $(h-1)$ non-zero edges can be connected to the selected node of one cluster, since based on pigeonhole principle, there would have to be at least one cluster with more than one selected node which violates constraint 1. Finally, since constraint 2 enforces the summation of edge values, which are smaller or equal to one, to be exactly $(h-1)$, and there cant be more than $(h-1)$ non-zero edges, thus the edge values have to be exactly 1. ■

Therefore, we need to enforce only the node variables to be binary and allow the edge variables to be continues, ranging from 0 to 1. We enforce this using $\mathbf{MX} \leq \mathbf{N}$, where $\mathbf{M}$ is defined as:

$$\{1 \leq i \leq \binom{n}{2}.k^2, 1 \leq j \leq (h.k + \binom{n}{2}.k^2)\} :$$
$$\mathbf{M}(i,j) = \begin{cases} 1 & \text{if } j = h.k + i, \\ 0 & \text{otherwise,} \end{cases} \quad (15)$$

and $\mathbf{N}$=1. Enforcing a variable to accept only binary values implies an additional constraint in an Integer Programming framework. Hence, by decreasing the number of binary variables, the complexity of our GMCP solver significantly reduces while the optimal solution is yet guaranteed.

The main contributing factors to the complexity of GMCP are the number of clusters and the number of nodes therein. We used Cplex [1] to solve our MBIP instances on a quad core 2.4 GHz machine. Utilizing the proposed solver, the GMCP instances corresponding to about 10,000 TRECVID-MED 2011-2012 videos could be solved in the negligible time of 2.04 seconds on average. In order to have a comparison with existing approximate GMCP solvers, we tested a method based on Local Neighborhood Search (LNS) [17, 6, 18] on the same instances. LNS and our method converged to the optimal answer in 83% and 100% of the cases respectively which confirms the optimality of our solver. The average search time by LNS was shorter (0.14 seconds); however, regarding the negligible time of solving a GMCP compared to the other required steps such as feature extraction, and codebook search, the solving time is not a crucial factor in our framework.

5. Experimental Results

Our method is suitable for *multiclass* classification, and not detection, since our class utility values (eq. 7) are meaningful on a comparative manner. Therefore, our method is unsuitable for solving the binary detection problem such as the one addressed in the TRECVID-MED task. However, even though we do *not* solve the TRECVID problem, we use the TRECVID-MED videos as our dataset. That is due to the complexity of such videos, their typical length and the intra-class variability which provide a good basis for evaluating a contextual co-occurrence based video classification method. Additionally, a broad range of annotations for semantic concepts are available for TRECVID-MED data which facilitates the evaluation process.

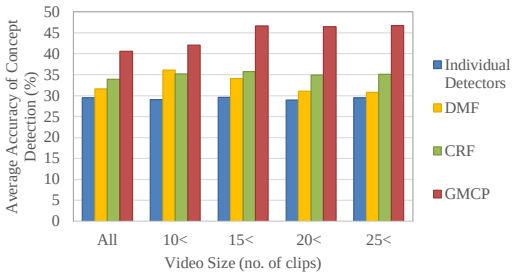


Figure 5. Performance evaluation of concept detection using the GMCP-based method and the baselines.

TRECVID11-MED and TRECVID12-MED [2] are currently among the most challenging datasets of complex events. We evaluate the proposed framework on EC11, EC12 and DEVT datasets. DEVT (8100 videos) is part of TRECVID-MED 2011 with of fifteen complex events of *Boarding trick, Feeding animal, Landing fish, Wedding, Wood working project, Birthday party, Changing tire, Flash mob, Vehicle unstuck, Grooming animal, Making sandwich, Parade, Parkour, Repairing appliance, and Sewing project*. EC11 and EC12 are subsets of TRECVID-MED 2011 and 2012 datasets and include 2,062 videos with annotated clips; in each video, the beginning and end of the video segments in which one of our 93 concepts occur are marked manually resulting in total number of 10,950 annotated clips. EC12 includes additional ten events of TRECVID-MED 2012. Note that the annotated clips (shots) are used only for training concept detectors and evaluating the concept detection results. The annotated clips in query videos are not used during test; we employ a sliding window approach for detecting the concepts in them (see sec. 5.2).

In order to train concept detectors, we extracted Motion Boundary Histogram (MBH) [14] features from the annotated clips and computed a histogram of visual words for each. Then, we trained 93 binary SVMs [4] with χ^2 kernel using the computed histogram of visual words.

5.1. Concept Detection Evaluation

We evaluated the proposed GMCP-based *concept detection* method on EC11 and EC12 using 10-fold cross validation scenario. We extracted the reference co-occurrence matrix, shown in fig. 3, utilizing the annotated clips of 9

folds and used the rest of the videos for testing.

As the first baseline, we applied the 93 individual SVM concept detectors to each annotated clip and picked the class with the highest confidence as the detected concept; this approach results in the average accuracy of 29%. In order to compute the accuracy, first we divide the number of correctly recognized clips by the total number of clips for each of the 93 concepts; the average of the resulting 93 values yields the average concept accuracy reported in fig. 5. The proposed GMCP-based concept detection method yields the average accuracy of 41%. Fig. 5 illustrates the accuracy of the baselines and the proposed method along with the breakdown with respect to the size of the test videos. As apparent in the chart, the improvement made by the proposed method increases as the size of the video grows, resulting in an improvement of over 16% for videos with more than 15 clips (about 1.6 minutes long). This is due to the fact that there are more clips in longer videos, and therefore, more contextual information to utilize; this shows that the proposed method is successfully leveraging the context.

We also provide the results of employing linear chain conditional random fields for concept detection, as CRF is a common approach to exploiting the context [13]. Additionally, fig. 5 shows the performance of using Discriminative Model Fusion (DMF) [7]; the negligible improvement made by DMF is consistent with observation made in other works [13], specially for large concept lexicons. Fig. 6 shows the confusion matrix of concept detection using the individual detectors and the proposed method; a notable improvement is observed on the main diagonal of GMCP results.

The improvement GMCP yields over CRF (and generally standard graphical models) is mainly due to not enforcing any specific structure on the graph, such as being acyclic, and preserving its generality. Moreover, our graph is complete; thus, a graphical model equivalent to our input would have to be complete and consequently contain lots of loops while common graphical models' inference methods, such as belief propagation, are known to have a degrading performance in the presence of loops [16], whereas our optimization method does not deteriorate with inclusion of loops in its input. Note that the concepts which are typically confused by GMCP, such as no. 34 *jumping over obstacles*, have too small scores from their SVM detectors, so they can not be completely fixed by incorporating the context.

5.2. Classification using GMCP

In this experiment, we evaluate the performance of the method described in section 3 where both concept detection and event classification are performed by GMCP. In order to keep the test and training set totally disjoint, we extracted the event-specific co-occurrence matrices, samples shown in fig. 4, from the annotations of EC11 clips, and used DEVT videos as the test set. We applied the SVM concept detectors to sliding windows of 180 frames (average size of clips in the annotated set) with displacement size of

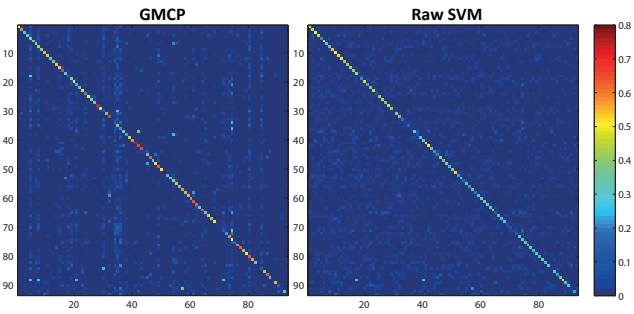


Figure 6. Confusion matrices of concept detection. Left and right show GMCP and the baseline results, respectively.

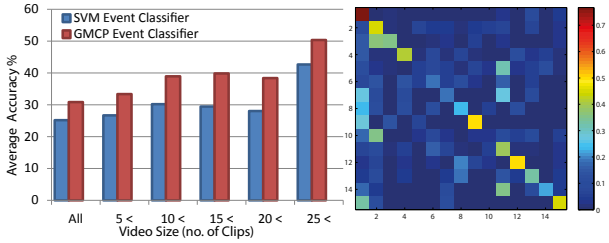
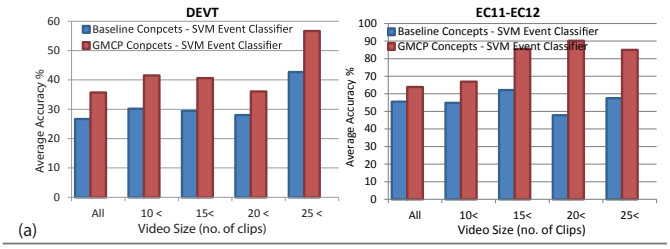


Figure 7. Performance evaluation of event classification using GMCP vs. SVM. Left and right show the bar chart of average accuracy and confusion matrix of GMCP respectively .

30 frames in each step. Therefore, each uniform clip of 180 frames has over 50% of overlap with six windows on which the SVM concept detectors were applied. We pick the window in which the highest SVM confidence value falls to represent the clip. We employ this approach since the beginning and end of the concept shots in a test video are unknown, and they are often overlapping. We ignore the clips for which the highest SVM confidence is less than 10%.

The video classification results of applying the GMCP-based classifier on these clips with the aforementioned concept detection scores are shown in Fig. 7-left. We use a *multiclass* SVM which performs the classification using the histogram of concepts occurring in the video as the baseline. This multiclass SVM is trained on the concept histograms extracted from the annotations of EC11. For the test videos, the concept with the highest score (by the individual concept detectors) is selected for each clip, and a histogram of all of the concepts found in the video is formed. This histogram is then classified using the multiclass SVM.

The bar chart in Fig. 7-left illustrates the average classification accuracy of GMCP and the baseline; the confusion matrix of GMCP results is shown on the right. As apparent in the bar chart, the proposed GMCP-based event classifier outperforms SVM, in particular for longer videos. This is consistent with the basis of the proposed approach as the event classification is being done in a contextual manner.



(a)

(b)

	All	10<	15<	20<	25<
GMCP-Concepts	**0.35**	0.35	**0.43**	**0.47**	**0.60**
Baseline-Concepts	**0.35**	**0.40**	0.41	0.38	0.39

Figure 8. Evaluation of event recognition using the concepts detected by GMCP and SVM. Part (a) compares the results on DEVT and EC11-EC12 datasets. Part (b) provides the mAP values.

5.3. Event Recognition using GMCP Concepts

We performed an experiment to verify how much improvement the concepts detected by GMCP cause in the overall performance of event recognition. We computed the histogram of 93 concepts for the annotated videos in EC11 and EC12 and used them for training fifteen binary classifiers representing fifteen TRECVID-MED11 events.

In one experiment, we used DEVT videos as the test set. As an standard baseline method, the histogram of the detected concepts by the raw SVMs in the test video was classified using the fifteen binary SVM event classifiers. We applied the GMCP-based concept detection method on each test video and formed the histograms using the improved concepts; the resulting histogram was classified using the trained SVM event classifiers. The bar chart of fig. 8 (a)-left compares the performance of the baseline and the proposed method. The GMCP concepts improve the overall performance of event recognition by 8% to 14% in terms of average event recognition accuracy. Fig. 8 (b) compares the results of the same experiment in terms of mAP (mean average precision). As apparent in the table, a notable improvement is seen as the size of the videos increases.

We performed a similar experiment on EC11 and EC12 datasets which include 25 events, using 10-fold cross validation scenario; the results are provided in fig. 8 (a)-right.

Note that in this experiment, the event recognition was always performed using SVM (unlike sec. 5.2) while the concepts were detected using GMCP or the baseline.

6. Conclusion

We proposed a contextual approach to complex video classification using generalized maximum clique graphs. We defined a co-occurrence model based on conditional probability, and proposed to represent an event using the co-occurrence of its concepts. Then, we classified a video based on matching its co-occurrence pattern, represented by a clique, to the class co-occurrence patters. We also developed a novel optimal solution for GMCP using Mixed Binary Integer Programming. The proposed approach opens new opportunities for further research in this direction, and the evaluations showed our method significantly outperforms well established baseline.

Acknowledgment: This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S.Government.

References

[1] IBM ILOG CPLEX Optimizer, www-01.ibm.com/software/integration/optimization/cplex-optimizer/. 4, 6

[2] TRECVID multimedia event detection track, http://www.nist.gov/itl/iad/mig/ med11.cfm, 2011-2012. 6

[3] E. Althaus, O. Kohlbacher, H. Lenhof, and P. Muller. A combinatorial approach to protein docking with flexible side chains, 2008. RECOMB. 4

[4] C. Chang and C. Lin. Libsvm: A library for support vector machines, 2011. ACM Transactions on Intelligent Systems and Technology 27:1–27:27. 6

[5] Z. W. et al. Youtubecat: Learning to categorize wild web videos, 2010. CVPR. 1

[6] C. Feremans, M. Labbe, and G. Laporte. Generalized network design problems, 2003. European Journal of Operational Research Volume 148, Issue 1. 3, 4, 6

[7] G. Iyengar and H. J. Nock. Discriminative model fusion for semantic concept detection and annotation in video, 2003. Proceedings of the eleventh ACM international conference on Multimedia. 2, 7

[8] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model, 2012. ECCV. 1

[9] Y. Jiang, J. Yang, C. Ngo, and A. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study, 2010. IEEE Trans. on Multimedia. 1

[10] A. Koster, S. V. Hoesel, and A. Kolen. The partial constraint satisfaction problem: Facets and lifting theorems, 1998. Operations Research Letters 23. 4

[11] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. S. Sawhney. Video event recognition using concept attributes. In *WACV*, pages 339–346, 2013. 2

[12] X. Liu and B. Huet. Automatic concept detector refinement for large-scale video semantic annotation, 2010. International Conference on Semantic Computing. 1

[13] A. L. w. Jiang, Shi-fu Chang. Context-based concept fusion with boosted conditional random fields, 2007. International Conf. on Acoustics, Speech and Signal Processing. 2, 7

[14] H. Wang, A. Klaser, C. Schmid, and C. L. Liu. Action recognition by dense trajectories, 2011. CVPR. 6

[15] X. Wei, Y. Jiang, and C. Ngo. Concept-driven multi-modality fusion for video search, 2011. IEEE Trans. on Circuits amd Systems for Video Technology. 1

[16] Y. Weiss. Correctness of local probability propagation in graphical models with loop. In *Neural Computation 2000*. 7

[17] A. Zamir and M. Shah. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2014. 4, 6

[18] A. R. Zamir, A. Dehghan, and M. Shah. GMCP-Tracker: global multi-object tracking using generalized minimum clique graphs, 2012. ECCV. 4, 6