Semi-supervised understanding of complex activities from temporal concepts

Carlos Fernando Crispim-Junior, Michal Koperski, Serhan Cosar, Francois Bremond

Semi-supervised understanding of complex activities from temporal concepts

Carlos Fernando Crispim-Junior, Michal Koperski, Serhan Cosar, Francois Bremond
Inria Sophia Antipolis
2004 Route des Lucioles, Valbonne, France
{carlos-fernando.crispim_junior, michal.koperski,francois.bremond}@inria.fr
scosar@lincoln.ac.uk

Abstract

Methods for action recognition have evolved considerably over the past years and can now automatically learn and recognize short term actions with satisfactory accuracy. Nonetheless, the recognition of complex activities - compositions of actions and scene objects - is still an open problem due to the complex temporal and composite structure of this category of events. Existing methods focus either on simple activities or oversimplify the modeling of complex activities by targeting only whole-part relations between its sub-parts (e.g., actions). In this paper, we propose a semi-supervised approach that learns complex activities from the temporal patterns of concept compositions (e.g., "slicing-tomato" before "pouring_into-pan"). We demonstrate that our method outperforms prior work in the task of automatic modeling and recognition of complex activities learned out of the interaction of 218 distinct concepts.

1. Introduction and Motivation

The task of automatic recognition of human activities has been studied for several decades [1][13][16][28], but it remains an open-problem due to the complexity of this category of events (*e.g.*, high-intra class variance complemented by low inter-class variance). A complex activity may be seen as a composition of two or more sub-parts, where a sub-part could be atomic (*e.g.*, actions and objects, namely a concept), or other activities. Moreover, an activity may be characterized by specific time arrangements between its sub-parts. For instance, during the activity of "serving a coffee", after the action of "pouring coffee into a cup", one may "put sugar into the cup" before or after "pouring the coffee", or even not put sugar at all. These three variations of "serving a coffee" are valid instances of the same activity.

To successfully recognize a complex activity (*e.g.*, serve and drink coffee), it is then necessary to model the temporal and composite relations between activities and their sub-parts (*e.g.*, "taking cup", "pouring coffee", "drinking coffee"), and when applicable, the semantic relations that permit to one or more of those sub-parts to be absent (*e.g.*, "pouring sugar", "pouring milk") when others are compulsory (*e.g.*, "taking cup", "pouring coffee").

Two main problems generally come up in computer vision approaches when modeling activities from relations among concepts. Firstly, the uncertainty due to unreliable observations from visual concept detectors, and secondly, the fact that the complexity of activity models tend to exponentially increases with the number of concepts and the order of their relations. For instance, consider a scenario where activities are composed of at most 10 concepts. The number of possible composite relations (pairwise combinations) between these concepts is 45 (C_2^{10}), a rather low value. However, the number of possible time arrangements (permutations - P) between these 45 composite relations is aproximately 2000 ($P_2^{C_2^{10}}$). Now, consider a more realistic number of concepts, *e.g.*, 218 (Cooking Composite data set [17]). The number of possible temporal relations between 218 concepts is approximately 559,440,756, a set of relations too large to be fully represented.

This paper proposes a semi-supervised, probabilistic framework to automatically learn complex activities from visual concept relations (Fig.3). We differ from previous methods [4][19][20][21][24][29] firstly by our deeper hierarchical representation of an activity (concepts, composite concepts, temporal composite concepts, and activities). Secondly, by the iterative, unsupervised framework we propose to find the temporal and composite relations between the concepts of targeted activities. The unsupervised learning of activity representations is complemented by a probabilistic model of characteristic frequencies of these relations.

The combination of these features make the propose approach uniquely capable of handling errors from underlying algorithms for concept recognition, like missed and/or overdetected concepts, scalable for large data sets (high-number of concepts and activities), and capable of describ-

ing an activity without human intervention.

The rest of this paper is organized as follows: section 2 presents the related work, section 3 introduces the proposed framework, section 4 describes the experimental procedures, and sections 5 and 6 present discussion and results and conclusions, respectively.

2. Related Work

Methods for complex activity recognition may be categorized into two groups: feature-based and structure-based methods. Feature-based methods are inspired on action recognition approaches, like [26][10], and focus on learning activity models directly from low-level data. To overcome the lack of semantics of this level of data, novel studies have separated activity modeling from raw pixel data by adopting an intermediate layer of atomic sub-parts, namely concepts [11][31][5][3][27][17][8][12]. Rohrbach *et al* [17] have decomposed each activity video into a set of temporal intervals, and then performed concept recognition per interval. To improve concept recognition they have reinforced concept scores of a temporal interval by considering patterns of concept co-occurrence learned either from training data or from text analysis of dish recipes. Activity recognition was performed using a nearest neighbor (NN) classifier with reinforced concept scores as input. In an alternative approach, Assari *et al.*[3] have converted the problem of relation discovery to a Generalized Maximum Clique Problem. Basically, they have searched over the time intervals of different videos for the set of concepts and relations that these video samples holistically agree on for each activity.

But, even though feature-based methods can easily scale for a large number of activity categories, most methods ignore temporal relations among concepts and formalize activities as patterns of absence / presence of concepts (*e.g.*, actions, objects). By ignoring such relations these approaches lose relevant information to discriminate among activity categories that share the same concept set, but display different temporal relations.

Alternatively, structure-based methods for activity recognition have focused on exploiting the hierarchical, composite nature of complex activities. Examples of such methods are probabilistic graphical models (PGM, *e.g.*, Hidden Markov Model, Dynamic Bayesian Networks)[30] and natural language processing techniques (NLP, *e.g.*, context free grammars and n-gram models) [7][24][9][25][15][18][6][23][22]. Siddharth *et al.* [18] have used regular expressions to model activity dependencies out of the possible roles that different concept categories may assume. Likewise, Vo et al. [25] have described the joint use of a stochastic CFG and a Bayesian network to represent the hierarchical structure of activities. They have shown that by operating on video time intervals, they could optimize activity recognition to perform exact inference.

In essence, structure-based approaches provide an optimal formalism to express the hierarchical and temporal dependencies between concepts and activities. On the other hand, they depend on the use of human knowledge to manually craft the rules among concepts (*e.g.*, relations, activity structure) that characterize each activity. By depending on human labor, these approaches become impractical as the number of activities and involved concepts increases. Moreover, since these methods are generally deterministic, their performance deteriorates in the presence of noisy observations.

In this paper, we propose a method that sits in between feature- and structure-based categories. We focus on the efficient, automatic discovery of concept relations for activity recognition. The proposed approach goes beyond previous methods by exploring short- to long-term temporal relations between the sub-parts of a complex activity. It resembles previous work of [4][19][20][21][24][29]. We differ from prior work in the following aspects: methods in [20] and [21] do not handle the uncertainty from automatically detected concepts, while the proposed method does. Models in [20][4][24][29] only target short-term relations between pairs of concepts. We model all possible temporal relations between pairs of concepts of any two time intervals of a video recording. Finally, we evaluate our method in the Cooking Composite data set [17], which contains 55 complex activities, composed of relations between 218 different concepts. Related approaches have evaluated their methods in much simpler and more constrained settings than ours (*e.g.*, 10 activities sub-divided into 10 sub-activities which define relations among 12 objects [21]).

3. Proposed method

The goal of the proposed approach is to automatically describe a video recording V given the temporal composite concepts observed during its time span. To accomplish this task we structure an activity model into four levels:

- **Concept** refers to a category of atomic observations of the scene, *i.e.*, a spoon, a person, the stirring action. It is equivalent to an attribute in [17], and an action in [3][9].
- **Composite concept**, or composites for short, corresponds to a pairwise relation between concepts within the same time interval.
- **Temporal composite (concept)** corresponds to a temporal relation between two composite concepts.
- **Activity** is a composition of temporal composite concepts.

We then refer to an instance as a particular observation (an example) of a concept, composite or temporal compos-

ite. Concept and composite instances are related to a time interval (t), while a temporal composite is related to two time intervals (t_a, t_b).

To handle a large set of concepts and their possible relations, we decompose the representation of a video into the following steps (see Fig.3): 1) video temporal segmentation, 2) concept recognition, 3) composite concepts and 4) temporal composite generation. Activity recognition step consists of finding the activity model, *a priori* learned, which most resembles the video representation of the analyzed video. This paper contribution focuses on the last two steps of this pipeline, which define our framework for the unsupervised learning of activity model based on concept relations, and a probabilistic method to perform activity recognition over these models.

Next subsections describe the pipeline of the proposed method: concept recognition (subsection 3.1), unsupervised video representation (subsection 3.2), and activity recognition (subsection 3.4). Activity model learning (subsection 3.3) takes place during a training phase over the unsupervised video representations of the training instances of each targeted activity.

3.1. Concept recognition

Given an activity video V split into $|T|$ time intervals (*e.g.,* a time interval is a video clip with homogeneous properties), concept recognition takes place at each time interval t and over the histogram of visual code words extracted from the corresponding time interval. Concept recognition is performed by a set of classifiers $\Psi = \{\psi_1, .., \psi_N\}$, one per concept category, that are trained *a priori* following a supervised learning scheme [17]. We employ a supervised learning scheme for concept recognition to reduce the amount of noise brought by this challenging task. The outcome of this step is a concept set $\Phi_1(t)$, which contains a concept instance ($\phi_{1,n}$) for each classifier active in the analyzed time interval (Eq.1). As a consequence, the number of concepts in a $\Phi_1(t)$ set varies between time intervals.

$$\Phi_1(t) = \bigcup_{n=1}^{N} \delta_n(t) \tag{1}$$

$$\delta_n(t) = \begin{cases} \phi_{1,n}, & \psi_n(t) = 1, \\ \emptyset, & \text{otherwise.} \end{cases}$$

where:

- $\phi_{1,n}$: concept instance from classifier n. Sub-index 1 denotes that this set element contains a single concept.
- N: number of concept classifiers,
- $\Phi_1(t)$: set of concept instances observed at time interval t,
- t: video time interval,
- δ_n: function that return concept instance ϕ_n given its observation by concept classifier ψ_n,
- $\psi_n(t)$: classifier that recognizes concept n applied to time interval t.

The notation $\Phi_1(T)$ defines the set which contains every concept set in video V, *i.e.*, $\Phi_1(T) = \{\Phi_1(1), ..., \Phi_1(|T|)\}$.

3.2. Unsupervised video representation

This step is responsible for constructing the unsupervised representation of an activity video V given the concepts observed in the $|T|$ time intervals of V. The output of this step is called activity video (or instance) representation. Its computation is sub-divided into two steps (Fig. 3, steps 3-4): generation of composite concepts and temporal composite concepts.

In essence, it starts by searching for composite concepts at each time interval, and only when this step is completed it moves to the analysis of the temporal relations between composite concepts. This iterative search makes activity inference more efficient, since it only considers concept patterns that are observed in the analyzed video.

3.2.1 Composite concepts

Composite concepts are pairwise associations (co-occurrences) of concepts within a given time interval. They may be seen as a special case of n-grams, in which concept (word) ordering is irrelevant during the analyzed time interval. To discover the composite concepts of a time interval (t), we compute the pairwise combinations of the L concept instances in its concept set, $\Phi_1(t)$ (Eq.2).

$$\Phi_2(t) = C_2^{|\Phi_1(t)|} = \{\phi_{2,l=0}, ..., \phi_{2,l=L}\} \tag{2}$$
$$\Phi_2(t=3) = \{\text{“}peel - apple'', ..., \text{“}hand - knife''\}$$

where,

- $\Phi_2(t)$: set of composite concepts of time interval t,
- $\phi_{2,l}(t)$: composite concept element l from time interval t,
- t: t-th time interval of the analyzed video.

3.2.2 Temporal composites

Temporal composites are discovered by analyzing the relations between the composite concept sets (time intervals) of an activity video. To obtain the temporal composite concepts between two time intervals $\in T$ (*e.g.*, t_1 and t_2),

we compute the Cartesian product between their composite concept sets (Eq. 3).

To represent an activity video we compute the temporal composite concepts between all pairs of time intervals in T of V, with no attention to the fact these intervals are consecutive or not (Eq. 4). For instance, given three time intervals, we generate temporal relations between elements of both consecutive ($\Phi_2(1), \Phi_2(1)$; $\Phi_2(2), \Phi_2(3)$) and non-consecutive composite sets ($\Phi_2(1), \Phi_2(3)$, see Fig.3). By making temporal composites invariant to the distance between their time intervals, the resulting activity model is more resilient to variations in the order that activity's subparts are carried-out.

$$\Gamma_2(t_1, t_2) = \Phi_2(t_1) \times \Phi_2(t_2) = \{ (\phi_{2,l}, \phi_{2,m}) \mid \phi_l \in \Phi_2(t_1) \; and \; \phi_m \in \Phi_2(t_2) \} \quad (3)$$

$$\Gamma_2(T) = \bigcup_{t_1, t_2 \in T} \Gamma_2(t_1, t_2) \quad (4)$$

where,

- $\Gamma_2(t_1, t_2)$: temporal composite concepts obtained from time intervals t_1 and t_2,
- Γ_2: set of all temporal composite concepts observed between the time intervals of a video.

Compared to Allen's interval algebra [2], the proposed method summarizes temporal relations among time intervals into "AND" and "BEFORE" relations. AND relation is modeled by composite concepts and catches temporal overlaps between concepts. BEFORE relation is modeled at the level of temporal composite concepts and it focuses on non-overlapping concepts. In the proposed representation other temporal relations (*e.g.*, OVERLAP, DURING, FINISH) are implicitly broken down and modeled in the context of "AND" and "BEFORE" relations.

3.3. Activity model learning

To learn the model of a given activity (*e.g.*, ω_i) the proposed method analyzes the concept relations shared among the learned representations of the J training instances of this activity (Eq.5).

$$\omega_i = \Gamma_2^{\omega_i} = \bigcap_{j=1}^{J(i)} \Gamma_2^j \quad (5)$$

$$\Omega = \{\omega_1, .., \omega_i, .., \omega_{|\Omega|}\} \quad (6)$$

where,

- Γ_2^j: set of temporal composite concepts of order 2 from video sample j,
- $\Gamma_2^{\omega_i}$: set of characteristic temporal composite concepts of activity ω_i at order 2,
- ω_i: activity model i,
- $J(i)$: J training samples of activity i,
- Ω: set of activity models.

Once the proposed method has identified the concepts and relations that characterize each activity of interest, it learns the expected frequency of these relations. For each temporal composite pattern $\gamma_{2,m}^{\omega_i} \in \Gamma_2^{\omega_i}$, we learn the Normal distribution, $\mathcal{N}(\mu, \sigma)$, that describes the frequency of $\gamma_{2,m}^{\omega_i}$ over the J training instances of activity i. As an outcome of this step, we obtain a probabilistic model about the temporal and composite relations among concepts of the targeted activity.

3.4. Probabilistic Activity Recognition

Given a set of activity models Ω (Eq.6) and a video representation (ω^*) constructed for a test video, activity recognition step consisting of computing the likelihood (Eq. 7) that ω^* belongs to one of the models in Ω.

$$P(\omega_*|\omega_i) = P(\Gamma_2^*|\Gamma_2^i) = \sum_{m=1}^{m=|\Gamma_2^i|} P(\gamma_m^*|\gamma_m^i) \quad (7)$$

The probability of each temporal composite concept is computed using Eq.8.

$$P(\gamma_m^*|\gamma_m^i) = \exp\Big(- \frac{(\, \theta(\gamma_m^*, T) - \mu_{\gamma_m^i})^2}{\sigma_{\gamma_m^i}^2} \Big) \quad (8)$$

where,

- θ: function that gives the frequency of γ_m^* in the T time intervals of the analyzed video recording,
- γ_m^*: temporal composite concept m of test video,
- $\mu_{\gamma_m^i}, \sigma_{\gamma_m^i}$: expected mean and standard deviation of the frequency of γ_m^i for activity i,

We assign to video V the label of the model ω_i that satisfies Eq. 9.

$$V = \underset{\Gamma_2^i}{\operatorname{argmax}} \, P(\Gamma_2^*|\Gamma_2^i) \quad (9)$$

4. Experiments

We evaluate the proposed approach on MPII Cooking Composite data set [17], a public data set that contains recordings of people cooking in a kitchen environment. The data set consists of 30 subjects performing 55 composite cooking activities (*e.g.*, how to prepare orange juice, a hot dog or coffee). This data set was chosen due to the large number of activities it contains, which are composed of an even larger number of distinct, low-level concepts (218). Subjects were free to perform the target cooking activity as they deem suitable, a factor which increases the difficulty for activity recognition. Each person is recorded by a color camera with a resolution of 1624×1224 pixels, on a frame-rate of 29.4 frames per second. The data set contains 256 videos in total and the duration of each video varies from 1 to 41 minutes. All videos are manually annotated for both cooking concepts and activities. To evaluate the performance of the proposed method we have followed the same train and test splits proposed by the data set authors. All reported results refer to test set recordings.

The data set divides each activity example into T time intervals. The size of T varies according to the number of steps (*e.g.*. $actions$) a given subject has performed to accomplish the targeted activity. We follow the time intervals and the method proposed by [17] for concept recognition. For instance, we analyze every time interval of a video sample for the presence of 218 concepts (*e.g.*. "pour", "put in", "egg", *etc*), using one SVM classifier per concept category in an one versus all scheme.

To obtain the concept observations of each time segment, we have proceeded as follows: first, we have ranked the 218 concepts by their classifiers' scores in the analyzed time interval, from the highest to the lowest. Then, one at a time we have tested the score of each classifier as a threshold for the 218 classifiers. Once we have determined the threshold with highest average f_1-score, we have used it to generate the concept instance set ($\Phi_1(t)$) of the time interval. The size of concept sets was limited to up the 10 most confident classifiers, as a measure to handle model complexity.

We have evaluated the proposed method with three experiments. First experiment verifies the contribution that pairwise temporal composites bring over a model using only composite concepts, either unary or pairwise relations. Second experiment compares the proposed method to prior work evaluated on MPII Cooking Composite data set. Finally, third experiment evaluates differences in performance between activity models using data-driven concepts (visual concept recognition) and concepts annotated by humans.

Performance is measured using weighted mean-average Precision (mAP) and f_1-score. Weighted scheme weights the performance of the evaluated method in each of the target classes by the class support in the test set [14].

5. Results and Discussion

The first experiment has evaluated the advantage of using temporal composite concepts (Γ_2) over using composite concepts (unary, Φ_1; pairwise,Φ_2). Results have shown that the use of pairwise temporal relations considerably improves activity recognition (59% mAP, see Table 1). It has also demonstrated that models using unary and pairwise composite relations contain relevant information for activity recognition. However, alone they are insufficient to achieve a competitive activity recognition rate.

Table 1. Cooking Composite Activity classification

Method	*Precision (%)*	f_1**-score**
Φ_1 model	15.74 ± 4.46	15.91 ± 5.20
Φ_2 model	18.38 ± 5.41	16.98 ± 4.80
Γ_2 model	59.42 ± 12.26	56.70 ± 10.00

The second experiments has compared the performance of the proposed method to prior work in MPII Cooking Composite data set (Fig.2). By exploiting temporal and composite relations between concepts, the proposed method has outperformed all baselines methods, even when they are using external sources of information (baseline *a*). We have observed that NN classifiers purely based on patterns of presence/absence of concepts (baseline b) have achieved a higher performance than both models using composite relations. This is most probably due to the fact NN classifiers implicitly learn the patterns of absence among the 218 concepts for each activity, an aspect that the proposed method does not exploit.

Finally, the third experiment has compared differences in performance between the proposed approach using annotated (GT) and automatically recognized concepts (data, Fig. 1). We have observed that the curves drawn by the precision and f_1-score values of activity recognition from ground-truth annotations have followed each other very closely. The same may be seen between the curves from automatically recognized concepts. This closeness between precision/f_1-score curves assures that the system recall is not sacrificed in detriment of achieving a higher precision. Nevertheless, we have observed a significant disparity in performance between models using data-driven and annotated concepts (approximately 20%). This gap points out that visual concept classifiers are still far from a satisfactory performance and, in the current state, they are one of the main limitations of the performance of the proposed method. Once these methods get more mature the proposed approach will naturally achieve its optimal performance.

6. Conclusions

In this paper, we have introduced a semi-supervised framework to learn complex activity models from tem-

poral and composite relations between concepts. We have demonstrated the proposed method surpasses baseline methods by modeling both short- and long-term relations among concepts. Its semi-supervised nature permits it to scale for large data sets, a setting in which structure-based approaches currently cannot. Finally, its iterative inference renders activity recognition tractable in the presence of a large quantity of concepts, a requirement that most existing methods cannot fulfill at the moment.

Further work will investigate methods to automatically segment a video recording into meaningful time intervals, more robust methods for concept recognition, and the relevance of higher order relations between concepts for activity understanding and recognition.

Acknowledgements

This research was supported by ANR SafEE project. We would like to thank authors of MPII Cooking Composite data set for their support during the usage of this data set, especially Dr. Marcus Rohrbach.

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, Apr. 2011.

[2] J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November 1983.

[3] S. Assari, A. Zamir, and M. Shah. Video classification using semantic concept co-occurrences. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2529–2536, June 2014.

[4] A. Behera, A. Cohn, and D. Hogg. Workflow activity monitoring using dynamics of pair-wise qualitative spatial relations. In K. Schoeffmann, B. Merialdo, A. Hauptmann, C.-W. Ngo, Y. Andreopoulos, and C. Breiteneder, editors, *Advances in Multimedia Modeling*, volume 7131 of *Lecture Notes in Computer Science*, pages 196–209. Springer Berlin Heidelberg, 2012.

[5] S. Bhattacharya, M. Kalayeh, R. Sukthankar, and M. Shah. Recognition of complex events: Exploiting temporal dynamics between underlying concepts. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2243–2250, June 2014.

[6] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman. Detection and explanation of anomalous activities: representing activities as bags of event n-grams. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 1031–1038 vol. 1, June 2005.

[7] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):852–872, Aug 2000.

[8] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, volume 7575 of *Lecture Notes in Computer Science*, pages 430–444. Springer Berlin Heidelberg, 2012.

[9] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[10] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, Sept. 2005.

[11] G. Luo, S. Yang, G. Tian, C. Yuan, W. Hu, and S. Maybank. Learning human actions by combining global dynamics and local appearance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(12):2466–2482, Dec 2014.

[12] F. Markatopoulou, V. Mezaris, N. Pittaras, and I. Patras. Local features and a two-layer stacking architecture for semantic concept detection in video. *IEEE Transactions on Emerging Topics in Computing*, 3(2):193–204, June 2015.

[13] T. B. Moeslund, A. Hilton, and V. Krger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(23):90 – 126, 2006. Special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[16] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010.

[17] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I*, pages 144–157, 2012.

[18] N. Siddharth, A. Barbu, and J. M. Siskind. Seeing what you're told: Sentence-guided activity recognition in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[19] M. Sridhar, A. G. Cohn, and D. C. Hogg. Learning functional object-categories from a relational spatio-temporal representation. In *Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 606–610, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.

[20] M. Sridhar, A. G. Cohn, and D. C. Hogg. Discovering an event taxonomy from video using qualitative spatio-temporal graphs. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 1103–1104, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press.

[21] J. Tayyub, A. Tavanai, Y. Gatsoulis, A. Cohn, and D. Hogg. Qualitative and quantitative spatio-temporal relations in daily living activity recognition. In D. Cremers, I. Reid, H. Saito, and M.-H. Yang, editors, *Computer Vision – ACCV 2014*, volume 9007 of *Lecture Notes in Computer Science*, pages 115–130. Springer International Publishing, 2015.

[22] C. Thurau. Behavior histograms for action recognition and human detection. In A. Elgammal, B. Rosenhahn, and R. Klette, editors, *Human Motion Understanding, Modeling, Capture and Animation*, volume 4814 of *Lecture Notes in Computer Science*, pages 299–312. Springer Berlin Heidelberg, 2007.

[23] C. Thurau and V. Hlav. n-grams of action primitives for recognizing human behavior. In W. Kropatsch, M. Kampel, and A. Hanbury, editors, *Computer Analysis of Images and Patterns*, volume 4673 of *Lecture Notes in Computer Science*, pages 93–100. Springer Berlin Heidelberg, 2007.

[24] K. Tu, M. Pavlovskaia, and S.-C. Zhu. Unsupervised structure learning of stochastic and-or grammars. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1322–1330. Curran Associates, Inc., 2013.

[25] N. N. Vo and A. F. Bobick. From stochastic grammar to bayes network: Probabilistic parsing of complex activity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[26] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[27] X. Wang and Q. Ji. A hierarchical context model for event recognition in surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2561–2568, June 2014.

[28] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224 – 241, 2011.

[29] C. Wu, J. Zhang, S. Savarese, and A. Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[30] H. Yu and J. M. Siskind. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 53–63, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[31] Y. Zhu, N. Nayak, and A. Roy-Chowdhury. Context-aware activity modeling using hierarchical conditional random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(7):1360–1372, July 2015.

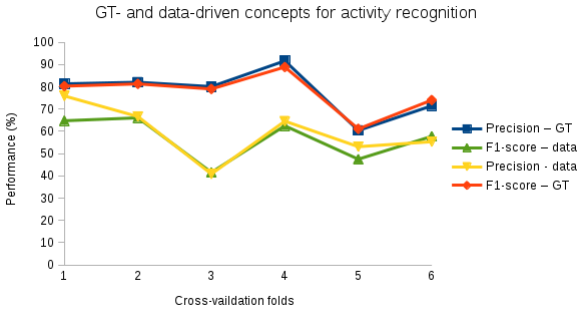

Figure 1. Complex activity recognition with annotated (blue and red lines) and automatically detected concepts (yellow and green lines) according to precision and f_1-score indexes.

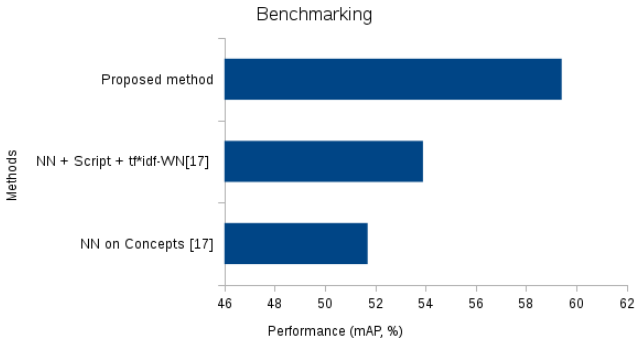

Figure 2. Performance benchmarking of our approach against data set baselines: *a*) Nearest Neighbor classifier (NN) on concepts, script data, and tf*idf-WN, and *b*) NN only on concepts.

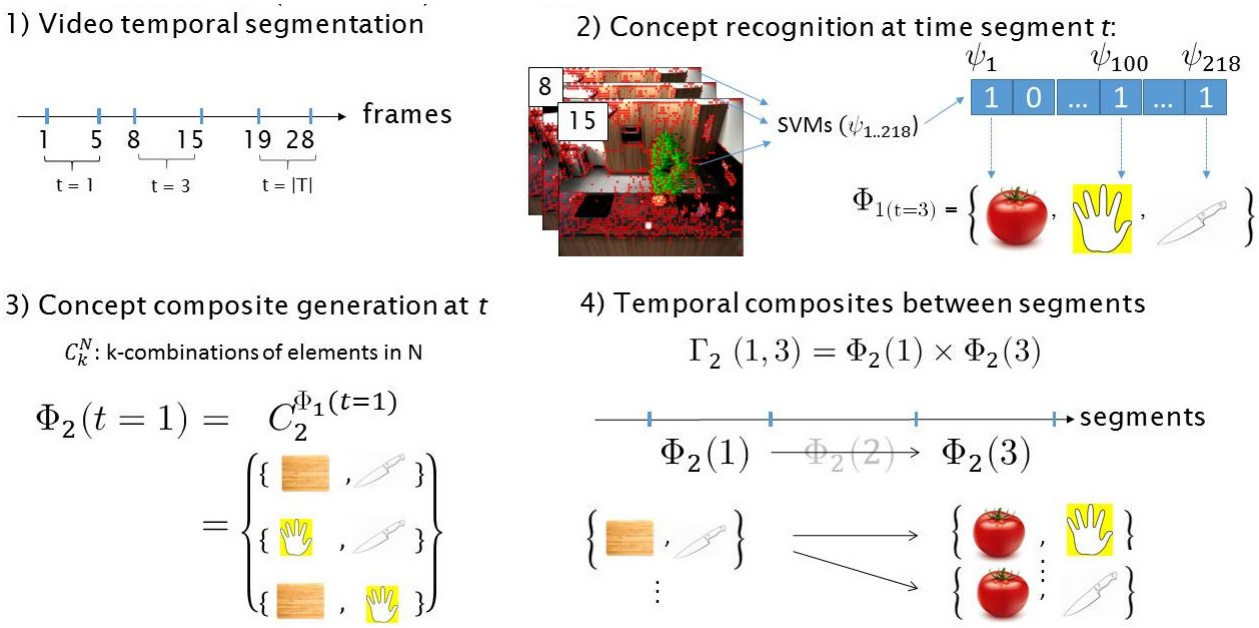

Figure 3. Semi-supervised learning of a video representation: 1) video temporal segmentation, 2) concept recognition 3) composite concept generation per time segment, 4) Temporal composite generation between segments.