Action Recognition in Videos

Christian Wolf and Atilla Baskurt
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR 5205, F-69621, France
e-mail: christian.wolf@liris.cnrs.fr, atilla.baskurt@liris.cnrs.fr

Abstract— **Applications such as video surveillance, robotics, source selection, and video indexing often require the recognition of actions based on the motion of different actors in a video. Certain applications may require assigning activities to several predefined classes, while others may rely on the detection of abnormal or infrequent activities. In this summary we provide a survey of dominant models and methods and discuss recent developments in this domain. We briefly describe two recent contributions: joint level feature and sequence learning, as well as space-time graph matching.**

Keywords— **Action recognition, motion, sequence modelling**

I. Introduction

Activity recognition requires robust motion estimation and spatio-temporal modelling of human motion, two research topics in their own right. The former consists in separating the video signal in an appearance component and a motion component. This is challenging in the case of moving cameras, when ego-motion needs to be separated from actor motion. The modelling step must takes into account temporal aspects and spatial aspects and their inter-dependence.

While early work on modelling human activities focused on articulated motion, most recent work does not explicitly model the human body[1]. Instead, the current state of the art focuses on the extraction of dense or sparse feature followed by learning or matching. Among the multiple possibilities for motion and feature extraction we cite the two dominant frameworks : (i) foreground/background segmentation, described in section II; (ii) sparse local features, described in section III. Alternatives, like dense matching and optical flow features, are beyond the scope of this summary.

II. Segmentation based methods

FG/BG segmentation produces a binary video stream which makes it possible to detect moving objects. Tracking methods can be easily integrated if required, for instance if person identification after occlusion is an issue. As a result, binary silhouettes are available for each moving person, thus shape based modelling is possible without further preprocessing. A frequent *Ansatz* is to model the evolution of shape (and other) features over time. Typically, a vectorial description is created frame by frame and an action is represented as a sequence of feature vectors, and learned with sequence models like HMMs [4], CRFs, recurrent networks [1] etc. Alternative methods consider an action as a binary three-dimensional object in space-time created, and characterize this 3D object through its volumetric properties or its projection onto 2D temporal templates. These methods are powerful, but they can properly work in controlled settings only, where background clutter, illumination variations, shadows, clothing and camera movement are not an issue.

Recent development tries to escape the dependence on fragile shape descriptors through automatic deep learning of features. In this context, and as an extension of work in object recognition [9], we proposed a translation variant sparse convolutional auto-encoder, which automatically learns a sparse feature extractor [1]. We were able to report excellent results for action recognition and facial expression recognition.

III. Keypoints and (semi)-structured models

Local features collected on sparse sets of points provide a compact yet rich representation, which is robust against occlusion and bypasses the tedious and error-prone segmentation task. The representation is inherently structural and is therefore difficult to use in a statistical learning framework. As a consequence, the set of local features is often converted into a numerical representation, discarding all or most of the structural information in the process. A typical example is the bag-of-words (BoW) formalism, originally developed for image classification and extended to action classification [7]. A visual dictionary is created through unsupervised learning and actions are represented as histograms over visual words.

A large variety of extensions to the BoW model have been proposed. We recently introduced a fully supervised way for jointly learning the dictionary and the prediction model, which makes the method more discriminant [6].

The biggest drawback of BoW models are their complete lack of spatial / spatio-temporal modeling, as all keypoint positions are discarded. This makes the model extremely invariant but severely limits discriminative power. Attempts have been made to extend the model to semi-structured models and up to to fully structured models like graphs. In figure 1 we illustrated the most widely used (semi)-structured models for action recognition in an attempt to put them on a linear scale between maximum invariance and maximum discriminance. Of course not all methods can be classified this clearly, but very often the two concepts *are* contradictory.

A powerful extension of BoW models, which significantly boosted discriminance with only a small decrease in invari-

[1] Articulated motion has again become widely used in the context of consumer depth cameras (Kinect). However, until now this has been restricted to actions in quite constrained environments without clutter.

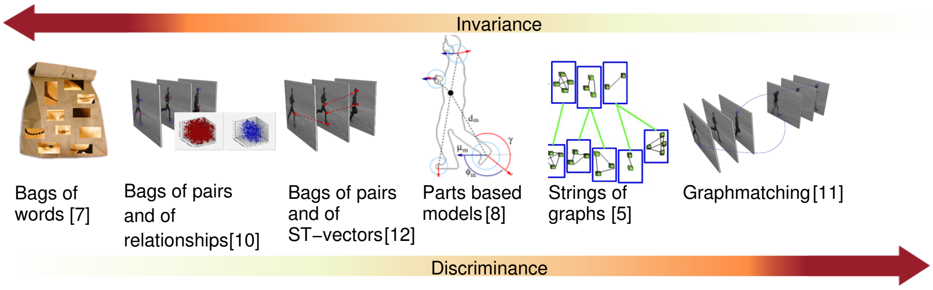

Fig. 1. Structured models and the trade-off between invariance and discriminance. The illustration for [5], [8] are reproduced from the respective papers.

ance, are pairwise models. In [10], pairs of keypoints are stored in 3D histograms, where the third dimension corresponds to a discretization of the spatio-temporal relationships. We extended this concept by adding a second dictionary which encodes the space-time vector between the two points of each pair [12]. Through cross-dataset learning, we showed, that this geometrical information is far robuster to changes in acquisition conditions than the appearance information alone, and in contrast to [10], this information can be used exclusively. Parts based models, initially developed for object recognition, have also been introduced successfully in action recognition. They decompose an action into local parts which are detected individually [8].

On the other end of the spectrum, graphs and hyper-graphs are fully structured models which form a natural description of this type of data. In [5] matching is done via temporally ordered local feature-graphs where each graph models spatial configuration of the features in a small temporal segment. In [11], we introduced a representation of actions as graphs built on ST keypoints and proximity information. Given a model keypoint set and a scene keypoint set, a possible solution of the problem is given through the values of a set variables x_i, where a value of $x_i{=}j$ is interpreted as model point i being assigned to scene point j. Matching is done minimizing a variant of a classical energy function known in object recognition [13]:

$$E(x) = \sum_i U(x_i) + \lambda \sum_{(i,j,k)\in\mathcal{E}} D(x_i,x_j,x_k) \qquad (1)$$

where U is a data attached term taking into account feature distances, D is the space-time geometric distortion between two triangles, $\mathcal{E}$ are the hyper-edges and λ_2 is a weight. Minimizing this kind of energy functions is in general NP-hard [13]. We recently showed [3], that the solution to this problem can be calculated in polynomial time in the case of keypoints embedded in space-time, due to three properties of the time dimension: (i) causality — actions cannot be reversed in time; (ii) uniqueness of time instants — points of the same frame are matched to a unique frame; (iii) limited time warping.

Our proposed algorithm is related to sequence alignment in that it exploits temporal information and its linear nature in a similar way. However, we do not perform simple sequence alignment. The novelty of our approach is that we use a full-fledged hyper-graph model with all its rich structural information stored in its nodes, embedded in space-time, and in its hyper-edges built from proximity information. The derived minimization algorithm is capable of dealing with classical energy functions including unary, binary and ternary terms, which makes it possible to include scale invariant potentials.

IV. The Future

Current challenges in this domain are complex activities,i.e. long duration, human-object interactions, complex human-human interaction etc., and context dependent modelling. We recently proposed a new dataset for these actions, which is also used for the ICPR 2012 HARL competition[2].

References

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Spatio-temporal convolutional sparse auto-encoder for sequence classification. In *Pr. of BMVC*, 2012.

[2] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *Pr. of ICPR*, 2011.

[3] O. Celiktutan, C. Wolf, and B. Sankur. Fast exact matching and correspondence with hyper-graphs on spatio-temporal data. TR RR-LIRIS-2012-002, INSA-Lyon, 2012.

[4] N.P. Cuntoor, B. Yegnanarayana, and R. Chellappa. Activity modeling using event probability sequences. *IEEE Transactions on image processing*, 17(4):594–607, 2008.

[5] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A "string of feature graphs" model for recognition of complex activities in natural videos. In *Pr. of ICCV*, 2011.

[6] M. Jiu, C. Wolf, C. Garcia, and A. Baskurt. Supervised learning and codebook optimization of bag of words models. *Cognitive Computation*, 2012.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Pr. of CVPR*, pages 1–8, 2008.

[8] K. Mikolajczyk and H. Uemura. Action recognition with appearance–motion features and fast search trees. *CVIU*, 115(3):426–438, 2011.

[9] M.A. Ranzato, F.J. Huang, Y.L. Boureau, and Y. Lecun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Pr. of CVPR*, 2007.

[10] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In *Pr. of ICCV*, 2009.

[11] A.-P. Ta, C. Wolf, G. Lavoue, and A. Baskurt. Recognizing and localizing individual activities through graph matching. In *Pr. of AVSS*, 2010.

[12] A.P. Ta, C. Wolf, G. Lavoué, A. Baskurt, and J-M. Jolion. Pairwise features for human action recognition. In *Pr. of ICPR*, 2010.

[13] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. In *Pr. of ECCV*, volume 2, pages 596–609, 2008.

[2] `http://liris.cnrs.fr/harl2012`