Joint Action Segmentation and Classification by an Extended Hidden Markov Model

Ehsan Zare Borzeshi, Oscar Perez Concha, Richard Yi Da Xu, and Massimo Piccardi

***Abstract*—Hidden Markov models (HMMs) provide joint segmentation and classification of sequential data by efficient inference algorithms and have therefore been employed in fields as diverse as speech recognition, document processing, and genomics. However, conventional HMMs do not suit action segmentation in video due to the nature of the measurements which are often irregular in space and time, high dimensional and affected by outliers. For this reason, in this paper we present a joint action segmentation and classification approach based on an extended model: the hidden Markov model for multiple, irregular observations (HMM-MIO). Experiments performed over a concatenated version of the popular KTH action dataset and the challenging CMU multi-modal activity dataset (CMU-MMAC) report accuracies comparable to or higher than those of a bag-of-features approach, showing the usefulness of improved sequential models for joint action segmentation and classification tasks.**

***Index Terms*—Action classification, action segmentation, Hidden Markov Model, joint segmentation and classification, probabilistic PCA, Student's t .**

I. Introduction and Related Work

AUTOMATIC recognition of human actions in video has been the focus of much recent research for its importance in applications such as video surveillance, human-computer interaction, multimedia and many others. Recognizing human actions is challenging since actions are complex patterns which evolve with time. Due to the human physiology, the way in which the same action is performed experiences major variations across subjects and instances. Moreover, a single subject often performs an action after another, and the problem of action recognition calls for the ability to jointly segment and classify the action sequence (a problem also known as sequential action labelling, or tagging).

Over the last decade, action recognition approaches have vastly leveraged on the notion of local spatio-temporal features [1]. Extracting such features consists of a detection and a description stages. The first stage requires detecting all the points in the actor's bounding box where significant "spatio-temporal change" occurs. The second stage consists of collecting a local descriptor for each detected point that summarizes its local spatio-temporal appearance. Actions as diverse as an elbow bending while picking a cup or a reclining head tend to generate specific descriptor values. As an example, Fig. 1 shows the points detected in the video of a person walking outdoor using the spatio-temporal interest points (STIPs) of Laptev *et al.* [2]. As shown in the sequence of frames, the detected points are irregular in both space (i.e. area in the frame) and time (i.e. number of points per frame). The same irregular nature in space and time is shared also by other, more specialized detectors such as the recently proposed poselet detectors [3]. In addition, descriptors are also typically high dimensional, affected by outliers and characterized by long-tailed statistics [4].

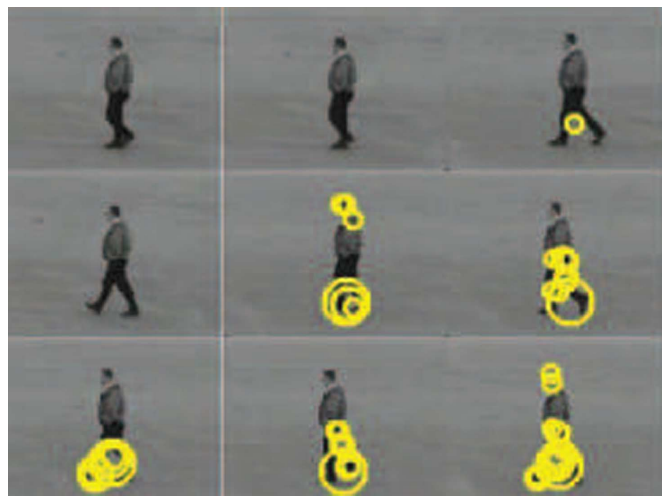

Fig. 1. Example of the spatio-temporal interest points from [2] in a video from the KTH action dataset. Frames are displayed in row-major order. The radius of circles is proportional to the scale at which change is detected. Note the variable number of points appearing in subsequent frames.

The baseline approach for action classification is known as "bag-of-features" [2], [5]–[7]. In this approach, the multi-dimensional descriptors are first quantized based on a learned codebook. Then, for each action instance, a histogram is computed over its quantized descriptors and used as input for a supervised classifier. Notwithstanding its simplicity, this approach has proved capable of remarkable recognition accuracy [2], [5]–[8]. Extension to segmentation can be obtained by simply splitting the video into overlapping windows and repeating classification for each window [9]. Yet, the size of the window and the overlap between windows are arbitrary,

Manuscript received June 01, 2013; revised July 24, 2013; accepted September 26, 2013. Date of publication October 01, 2013; date of current version October 14, 2013. The associate editor coordinating the review of this manuscript and aproving it for publication was Prof. Jeronimo Arenas-Garcia.

E. Zare Borzeshi, R. Y. D. Xu and M. Piccardi are with the University of Technology, Sydney, Australia (e-mail: Ehsan.ZareBorzeshi@uts.edu.au; YiDa.Xu@uts.edu.au; Massimo.Piccardi@uts.edu.au).

O. Perez Concha is with the Centre for Health Informatics, The University of New South Wales, Sydney, Australia (e-mail: O.PerezConcha@unsw.edu.au).

Supplemental multimedia files for this paper are available online at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2013.2284196

with possible impact on accuracy and temporal resolution. Temporal graphical models offer a more principled approach to the segmentation problem [10]–[12]. For this reason, in this paper we adopt an extended hidden Markov model - named hidden Markov model for multiple, irregular observations (HMM-MIO) hereafter - capable of providing classification and time segmentation over a) observations which are irregular in time and space, b) high-dimensional observation spaces, and c) outliers and heavy-tailed distributions. This model was recently proposed in [13] and is extended here over spatial regions. Experiments are performed over a "stitched" version of the popular KTH dataset [5] where individual actions have been collated into uninterrupted sequences, and the challenging CMU multi-modal activity dataset (CMU-MMAC) which displays scenes of cooking actions [14]. The achieved accuracies show that HMM-MIO is capable of competitive performance in joint action segmentation and classification.

The rest of the paper is organized as follows: in Section II, we describe the generative model of the extended HMM. In Section III, we present an experimental evaluation of the proposed method on the concatenated KTH and CMU-MMAC datasets. The Conclusion section summarizes the main contributions of this letter.

II. Hidden Markov Models for Action Recognition

The hidden Markov model (HMM) is a factorized model for the joint probability of a sequence of observations, $x_{1:T} \equiv \{x_1, \dots x_t, \dots x_T\}$, and a sequence of corresponding hidden states, $y_{1:T} \equiv \{y_1, \dots y_t, \dots y_T\}$. Under the well-known Markov and observation independence assumptions, the joint probability factorizes as $p(x_{1:T}, y_{1:T}) = p(y_1)\prod_{t=2}^{T} p(y_t|y_{t-1})\prod_{t=1}^{T} p(x_t|y_t)$. The hidden Markov model enjoys efficient exact solutions for canonical problems such as likelihood evaluation, state decoding and maximum-likelihood estimation under supervised states. Maximum-likelihood estimation with unsupervised states is a non-convex problem for which local optima can be found by expectation-maximization algorithms [15]. HMM offers a natural model for joint action classification and segmentation: each state y_t, $t = 1 \dots T$, is assumed to correspond to the action at that time frame. Given an observation sequence $x_{1:T}$, state decoding, i.e. $\hat{y}_{1:T} = \text{argmax}_{y_{1:T}} p(y_{1:T}|x_{1:T})$, retrieves the most probable action sequence. Model estimation is performed with supervised states using an annotated training set of action sequences.

A. HMM-MIO

In action recognition, typical local features are irregular in space and time and characterized by high dimensionality. In addition, their empirical distributions tend to exhibit heavy tails and outliers [4]. In [13], the authors proposed a model (HMM-MIO) where:

- Observations can be one, none or more than one per frame.
- High dimensionality is mollified by adopting the probabilistic principal component framework [16]. With this approach, the covariance matrix of each observation density, Σ, is constrained to decompose as $W^T W + \sigma^2 \mathbb{I}$, where W is a matrix of limited vertical size. This constraint equates to modelling the observations over a lower dimensional space with spherical noise.
- Both heavy-tailed statistics and outliers are taken into account by modelling the observation densities with a long-tailed distribution, the Student's t [17].
- Possible multimodality of the observation densities is accounted for by using mixture distributions.

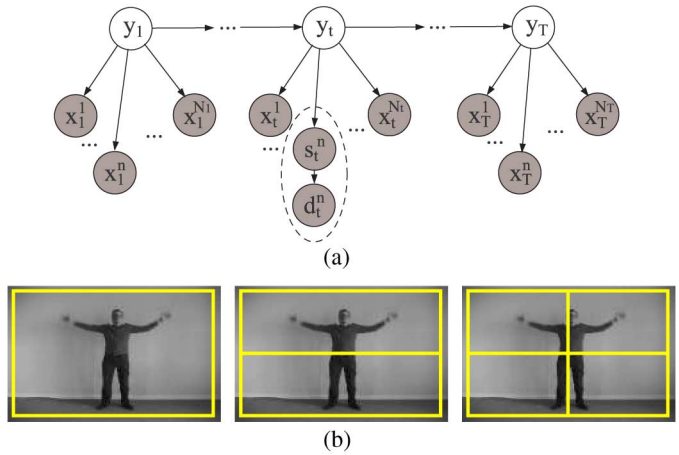

Fig. 2. a) The graphical model of HMM-MIO; b) the uniform grid over the actor's area.

In this letter, we add dealing with space irregularity by partitioning the bounding box containing the actor over a uniform grid with a small number of cells (typically, 1 to 4; see Fig. 2(b)), and using a separate density for the observations of each cell. The cell index, s, is a fully observed variable for each observation. We note each observation as x_t^n, with t the frame index and $n = 1 \dots N_t$ the observation index within the frame. Each observation consists of the pair, $x_t^n = \{d_t^n, s_t^n\}$, of the descriptor, d_t^n, and the cell index where it occurs, s_t^n. The state-conditional observation probabilities are assumed to factorize as $p(x_t^n|y_t) = p(d_t^n|s_t^n, y_t)p(s_t^n|y_t)$. For the multiple observations of a frame, we posit:

$$p(x_t^{1:N_t}|y_t) \equiv p(x_t^1, \dots x_t^{N_t}|y_t) = \prod_{n=1}^{N_t} p(x_t^n|y_t), \text{if } N_t \geq 1$$
$$= 1, \text{if } N_t = 0 \quad (1)$$

Posing $p(x_t^{1:N_t}|y_t) = 1$ in the case of no observations is equivalent to a missing observation and has neutral effect in the chain evaluation of the HMM. The generative model of HMM-MIO:

$$p\left(x_{1:T}, y_{1:T}\right) \equiv p\left(\{x_t^{1:N_t}, y_t\}_{t=1}^{T}\right)$$
$$= p(x_1^{1:N_1}, y_1, \dots x_t^{1:N_t}, y_t, \dots x_T^{1:N_T}, y_T) \quad (2)$$

is shown in Fig. 2(a).

Detailed evaluation, decoding and estimation formulas can be found in [13].

B. Comparison With Discriminative Sequential Models

In recent years, linear-chain conditional random fields (CRFs) have gained attention as an alternative to hidden Markov models [18]. The main advantage offered by CRFs

is their discriminative training, either as a probabilistic model or in a maximum-margin framework [19]. Their accuracy has been repeatedly reported as higher than that of corresponding HMMs (e.g., [10], [20]). However, there are two standing limitations which prevent extending a conditional random field with the features of HMM-MIO. The first limitation is that a principal component framework requires a log-quadratic model (for terms of the form $w_i w_j x_i x_j$) for which standard estimation algorithms are unsuited. The second limitation is the short tails of the exponential family on which CRFs are based. Conversely, the density of the Student's t is not exponential and enjoys an asymptotic value of $O(x^{-\nu-1})$ that can be modulated by the degree of freedom parameter, ν, to properly account for long tails and outliers. These considerations explain why a generative model like HMM-MIO offers complementary advantages to CRF for action recognition from local features.

III. Experimental Results and Discussion

A. Stitched KTH Dataset

For a first experiment on joint segmentation and classification, we have created a "stitched" version of the well-known KTH dataset by simply concatenating individual action instances into sequences [1]. Each sequence depicts a single actor in a homogeneous scenario (indoor, outdoor etc) performing a succession of 24 action instances for a total duration of approximately 2,000 frames. The actions were picked randomly, alternating between the two groups of {walking, jogging, running} and {boxing, hand-waving and hand-clapping} to emphasize action boundaries. A total of 64 such sequences were used for training and 36 for testing. The parameters selected over the training set were used unchanged for the test set.

Comparative experiments have been performed using HMM-MIO, classification of single frames and a bag-of-features approach. The number of reduced dimensions, D, and the number of components in each observation mixture, M, were made vary over interval (3,30). The degrees of freedom of the t distribution, ν, were made vary over {3,6,9} and the number of cells, S, over {1,2,4}. To implement single-frame classification, we have used a version of HMM-MIO with uniform transition probabilities. This equates to classifying frames solely based on the observation model, ignoring sequentiality in decoding. Frames with no observations were arbitrarily assigned to the first class in appearance order. For bag-of-features, we have used k-means clustering with $N = \{128, 256, 512\}$ clusters for quantization and an SVM classifier with RBF kernel for classification. In the test sequences, each window of $W = 32$ frames has been assigned a single action label, sliding the window forward one frame at a time. As features, we have extracted STIPs with the public software from [2], with the default descriptors of 162 dimensions each.

Table I shows the results on the test set in terms of frame-based accuracy, using the parameters' values that scored the best accuracy on the training set. The highest accuracy for the three compared models was achieved by HMM-MIO (71.2%). The importance of using a sequential model for segmentation is evidenced by the comparison with single-frame classification: the drop in accuracy is almost 30 percentage points. This drop is caused by both the arbitrary classification of frames without observations and the dismissal of the sequential context. The accuracy achieved by bag-of-features (61.8%) proved more than 9 percentage points lower than that achieved by HMM-MIO. The sensitivity to the parameters' values is not very pronounced: the range of accuracies for HMM-MIO is $\{66.5\% - 71.2\%\}$, $\{38.8\% - 41.8\%\}$ for single-frame classification, and $\{55.3\% - 61.8\%\}$ for bag-of-features.

TABLE I
Frame-Based Accuracy (%) for Joint Classification and Segmentation Over a Stitched Version of the KTH Dataset. S: Number of Cells; D: Number of Reduced Dimensions, M: Number of Components Per Mixture, ν: Degrees of Freedom; N: Number of Clusters; W: Window Size

Method	Parameters	Test accuracy (%)
HMM-MIO	$D = 30$, $M = 18$, $\nu = 3$, $S = 2$	71.2
Single-frame classification	$D = 30$, $M = 18$, $\nu = 3$, $S = 2$	41.8
Bag-of-features	$N = 256$, $W = 32$	61.8

B. CMU-MMAC Dataset

For a more probing and realistic experiment, we have tested our approach on a subset of the CMU Multi-Modal Activity Database (CMU-MMAC) containing multimodal measurements of the activity of forty subjects cooking following five different recipes [14]. For this experiment, we have selected the video clips of twelve different subjects making "brownies" from a dry mix box. The subjects attended to the preparation in a spontaneous way, without receiving instructions on how to perform each task; therefore, the action instances vary greatly in time span and manner of execution. Each video depicts a person performing a sequence of actions, with each action belonging to one of 14 classes including pouring, spraying, stirring, and others (see Fig. 3 for the complete list). The average duration of a video is approximately 15,000 frames while the average length of an action instance is approximately 230 frames, with a minimum length of 3 frames and a maximum of 3,269. As video source, we have used the view from static camera "7151062" which offers a side view of the scene (see Fig. 3). As action labels, we have used the annotations provided for the wearable camera mounted atop the subject's head, albeit only loosely synchronized with the static camera. For the experiment, we have used 12-fold cross-validation with a validation set, selecting eight subjects for training, three for validation and one for testing in each fold on a rotating basis. As features, we have again extracted STIPs with the public software from [2], but sub-sampling them one in ten in appearance order so as to limit the overall data size. The compared algorithms include HMM-MIO, single-frame classification, and the bag-of-features approach. For bag-of-features, we have extended the parameter search to {128,256,512,1024} for the number of clusters and {16,32,64} for the window size.

Table II shows the results from this experiment in terms of average accuracy and standard deviation over the folds. As it is far more realistic and challenging, accuracies are generally much lower. The best average accuracy is achieved by HMM-MIO

[1]All code and directions to reproduce our experiments are provided as Supplementary Material.

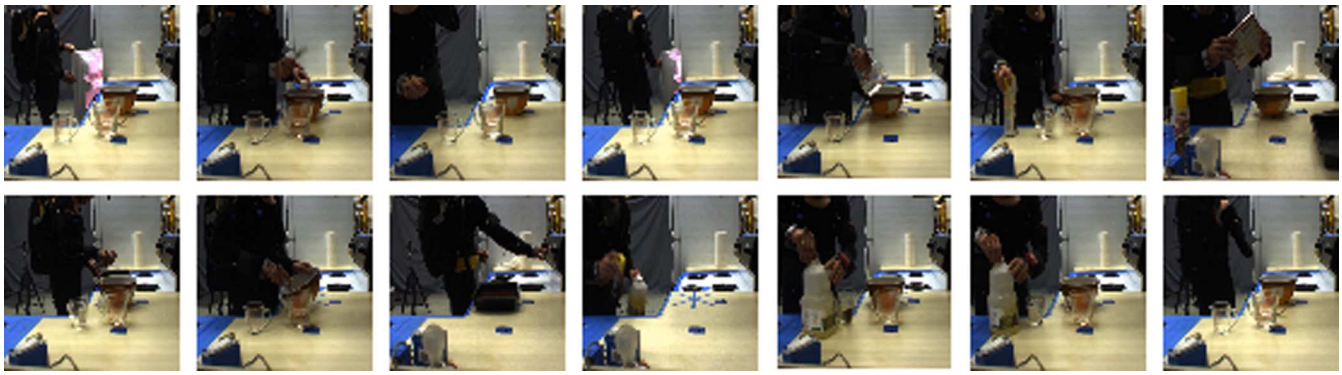

Fig. 3. Examples of actions for preparation of "brownies": (from left to right, column wise) *close, crack, none, open, pour, put, read, spray, stir, switch-on, take, twist-off, twist-on and walk*.

TABLE II
FRAME-BASED ACCURACY (%) FOR JOINT CLASSIFICATION AND SEGMENTATION OVER A SUBSET OF THE CMU-MMAC DATASET

Method	Average accuracy	Standard deviation
HMM-MIO	38.4	6.1
Single-frame classification	11.7	1.6
Bag-of-features	35.2	2.3

(38.4%) and is noticeably higher than that of bag-of-features (35.2%). However, HMM-MIO is more sensitive to the training fold as it reports a much higher standard deviation (6.1 vs. 2.3). The drop in accuracy with single-frame classification (minus 26.7 percentage points from HMM-MIO) is proportionally even more remarked than in the previous experiment, giving evidence to the importance of the sequential structure at a parity of observation model.

IV. CONCLUSION

In this letter, we have presented an approach to joint action segmentation and classification based on an extended HMM capable of exploiting local spatio-temporal features. Such measurements are irregular in space and time, high dimensional and characterized by heavy-tailed distributions and outliers. The extended model, HMM-MIO (hidden Markov model with multiple, irregular observations), effectively tackles these issues and provides significant accuracy. The accuracy reported over a stitched version of the KTH action dataset and the CMU-MMAC dataset proved comparable to or higher than that of a bag-of-features approach (71.2% vs. 61.8% and 38.4% vs. 35.2%, respectively) and much higher than that of single-frame classification with the same model (29.4 and 26.7 percentage points, respectively). These results show that sequential generative classifiers can be capable of significant action recognition accuracy, provided they are endowed with likelihood models that are well suited to typical visual measurements.

REFERENCES

[1] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. ICCV*, 2003, pp. 432–439.
[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. CVPR*, 2008, pp. 1–8.
[3] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proc. CVPR*, 2011, pp. 3177–3184.
[4] Y. Jia and T. Darrell, "Heavy-tailed distances for gradient based image descriptors," in *Proc. NIPS*, 2011, pp. 397–405.
[5] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proc. ICPR*, 2004, pp. 32–36.
[6] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *Proc. ICCV*, 2009, pp. 925–931.
[7] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. CVPR*, 2010, pp. 2046–2053.
[8] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009, p. 127.
[9] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic annotation of human actions in video," in *Proc. ICCV*, 2009, pp. 1491–1498.
[10] D. L. Vail, M. M. Veloso, and J. D. Lafferty, "Conditional random fields for activity recognition," in *Proc. AAMAS '07*, 2007, pp. 235:1–235:8.
[11] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *IEEE T-CSVT*, vol. 18, pp. 1499–1510, 2008.
[12] M. Hoai, Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *Proc. CVPR*, 2011.
[13] O. Concha, R. Da Xu, Z. Moghaddam, and M. Piccardi, "HMM-MIO: An enhanced hidden Markov model for action recognition," in *Proc. CVPRW*, 2011, pp. 62–69.
[14] F. D. la Torre, J. Hodgins, J. Montano, and S. Valcarcel, Detailed human data acquisition of kitchen activities: the cmu-multimodal activity database (cmu-mmac) Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. RI-TR-08-22h, 2008.
[15] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.
[16] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 61, no. 3, pp. 611–622, 1999.
[17] S. Chatzis, D. Kosmopoulos, and T. Varvarigou, "Robust sequential data modeling using an outlier tolerant hidden markov model," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31, pp. 1657–1669, 2009.
[18] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001, pp. 282–289.
[19] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, 2005.
[20] S. B. Wang, A. Quattoni, L.-P. Morency, and D. Demirdjian, "Hidden conditional random fields for gesture recognition," in *Proc. CVPR*, 2006, pp. 2:1521–2:1527.