Compositional behavior of HMMs in action recognition

F. Cuzzolin, R. Frezza

Dipartimento di Elettronica e Informatica
Università di Padova
35131 Padova, Italy

A. Bissacco, S. Soatto

Department of Computer Science
University of California at Los Angeles
90024 Los Angeles, CA

Abstract

In this paper we describe a characterization of "visual action" that encodes local photometry via a choice of interest operators and global dynamics via a realization of a stochastic dynamical model. In order to allow action detection in clutter, it is necessary for the corresponding models to have a compositional property, in that a simple action (e.g. foreground action) can be detected within a more complex one (e.g. foreground and background actions). We show that this is the case for the model we propose, which can therefore be used as a basis for building models of dynamic scenes from images without explicit supervision, by composing a complex action from a collection of elementary ones.

1 Introduction

Interpreting dynamic scenes remains one of the most important and yet largely untapped problems in computer vision, mainly because of the large variability in visual measurements of a same movement. A model of an "action" must embed invariance to all distracting factors either explicitly (i.e. as part of the model) or implicitly (by means of training).

For highly structured objects (like human bodies) detecting an action (e.g. a person crossing the road) can be subsumed into a parametric statistical estimation problem (see for instance [2, 11] and references therein). Invariance to some of the parameters (e.g. photometry) can instead achieved by choice of local features (e.g. optical flow), while a small number of parameters (e.g. lengths and angles between limbs) can be inferred from the data.

In this paper we present a general model of an "action" that includes both the dynamical and photometric characterization of the motion. We show that this model has *compositional* properties, in the sense that a model of a simple action can be detected from the model of a more complex one (in a dynamic "foreground + clutter" scenario). We adopt hidden Markov models as a method for extracting the invariant dynamic pattern from a collection of instances of the action. Its compositional behavior is tested by showing how the model of a simple action can be recovered from cluttered sequences.

In perspective, our long-term goal is the construction of models of spatio-temporal visual scenes from data without direct supervision, integrating top-down information, in the form of prior explicit models, with data driven bottom-up representations.

1.1 Related work

The problem of recognizing complex motion patterns in image sequences has been investigated in various settings. A common approach consists of extracting low-level features by local spatio-temporal filtering on the images and using hidden Markov models (HMMs) on the collection of sequences of points thus obtained for recognition and classification tasks [8]. In [10], parametric HMMs are introduced for recognizing gestures that exhibit dependence on a set of parameters, while in [1] coupled HMMs are used for modeling interactions of two mobile objects. Furthermore, in [4] more complex actions are recognized by computing the probability of a sequence of elementary actions detected by HMMs with a stochastic probabilistic parsing algorithm.

2 Modeling actions

A satisfactory definition of the concept of "action" requires an accurate analysis of the problem. The meaning of a motion is often *not* associated to the entire trajectory. For instance, in a pointing gesture the crucial part is the arm reaching a steady, extended position no matter what trajectory it follows. Trying to parameterize the nuisance variability of a class of trajectories leads to fragile models.

Even more important, the meaning of the action can be associated to *different* aspects of the movement: periodicity (closing the hand just once does not mean "bye!"), shape variations (e.g. hand gestures), critical configurations (the pointing gesture again), spatial trajectories (e.g. drawing). Therefore, in an unsupervised framework a multiple feature representation is necessary.

Consider a sequence of images $\{I_{t_1}, I_{t_2}, \ldots, I_{t_k}, \ldots\}$ and let $\{\phi_i, i = 1 \ldots P\}$ be a number of spatial or spatio-temporal filters, i.e. functions $\phi_i : \Omega_i \to \mathbb{R}^{k_i}$ where Ω_i is a subset of the image lattice. The location of the maximal response of each filter changes, in general, during the sequence, yielding a set of trajectories $\{f_i(t),\ i = 1 \ldots F, t = 1, 2, \ldots\}$, which we call *feature trajectories*. Only some of these feature trajectories will be of interest; we call the others the "background scene". Of course portions of feature trajectories may be missing.

We assume that feature trajectories of interest describe smooth transitions between a certain number N of configurations of maximal responses of filters $\{\nu_\tau,\ \tau = 1 \ldots N\}$, which we call *feature configurations*. They represent an invariant pattern that can be extracted from every instance of the action and does *not* necessarily encode the whole feature trajectories.
It is pretty natural to adopt the formalism of the *hidden Markov models* to represent this invariant pattern.

3 Learning actions

A *hidden Markov model* is a statistical model in which the states $\{X_k\}$ form a *Markov chain*; the only observable quantity is a corrupted version y_k of the state called *observation process*. Using the notation in [3] we can associate the elements of the finite state space $\mathcal{X} = \{1, ..., n\}$ to coordinate versors $e_i = (0, .., 0, 1, 0, .., 0) \in \mathbb{R}^n$ and write the model as

$$\begin{cases} X_{k+1} = AX_k + V_{k+1} \\ y_{k+1} = CX_k + diag(W_{k+1})\Sigma X_k \end{cases}$$

where $\{V_{k+1}\}$ is a sequence of martingale increments and $\{W_{k+1}\}$ is a sequence of i.i.d. Gaussian noises $\mathcal{N}(0, 1)$. The HMM parameters will be the *transition matrix* $A = (a_{ij}) = P(X_{k+1} = e_i | X_k = e_j)$, the matrix C of the *means of the state-output distributions* (in fact $C_j = E[p(Y_{k+1}|X_k = e_j)]$) and the matrix Σ of the variances of the output distributions.

A fundamental property of this class of models is the capability of self-learning the set of parameters A, C and Σ given a sequence of observations that are supposed to be produced by the system. The algorithm we use is an application of the EM technique [3]. The probabilistic distance between the measurement and each state representative $C \cdot e_j$ in the d-dimensional observation space

$$\Gamma^i(y_{k+1}) \doteq \prod_{j=1}^{d} \frac{g(\frac{y^j_{k+1} - c^j_i}{\sigma^j_i})}{\sigma^j_i g(y^j_{k+1})} e_i, \quad g \sim \mathcal{N}(0, 1)$$

is the error which is fed-back to drive the recursive estimator for the state:

$$\hat{X}_{k+1} = \sum_{i=1}^{n} A_i \langle \hat{X}_k, \Gamma^i(y_{k+1}) \rangle$$

where n is the number of states, A_i is the i-th column of A and $\langle \cdot, \cdot \rangle$ is the usual scalar product.

Assuming that the "correct" filters are given, HMMs provide a method to build the invariant pattern or graph between feature configurations. Feature configurations ν_τ are encoded as columns C_τ of the C matrix while their dynamics is associated to the transition probabilities in the A matrix.

4 Detecting actions in clutter

We want to test the compositional property of the model we described above, by showing that it is possible to recover the invariant pattern from the model of a complex motion. In other words, there is a map between the model representing a sequence containing both foreground and background motion and the model of the action of interest. We need to make the nature of this map precise, find a way of extracting the invariant information from an HMM built from a cluttered sequence and a criterion to compare the extracted information with the learned model of the action of interest.

Now let us suppose that we are given the suitable family of local features. It is reasonable to claim that *each state of the model generated in presence of clutter corresponds to a state of the learned model for the action of interest.*
Hence, the set of states of the first model associated to the same state of the learned HMM will roughly represent the same positions of the foreground features. Therefore, if we select the components of the state-output matrix C associated to the local features describing the action, the columns of the resulting matrix must form *clusters* in the associated subspace of the feature space. This operation can be done using a standard technique, for example *k-means clustering*, by considering the means of the state-output distributions collected in C as vectors in the d-dimensional feature space.

Once produced the set of n_c state clusters $\mathcal{C}^k = \{e^k_1, ..., e^k_{n_k}\}$ we need to rearrange the transition matrix in order to produce a new admissible model. This can be done considering one cluster at a time with no particular ordering, and grouping the corresponding

states.
After simple calculations we get

$$P(X_{t+1} \in \mathcal{C}^k | X_t = e_i) = \sum_{e_j \in \mathcal{C}^k} P(X_{t+1} = e_j | X_t = e_i)$$

while the transition probabilities *from* the cluster must be normalized with its cardinality:

$$P(X_{t+1} = e_i | X_t \in \mathcal{C}^k) = \frac{\sum_{e_j \in \mathcal{C}^k} P(X_{t+1} = e_i | X_t = e_j)}{|\mathcal{C}^k|}.$$

This operation is repeated for the next cluster on the new A until we eventually get an $n_c \times n_c$ matrix. Finally the columns of the reduced A must be permuted to match the order of the clusters in the reduced C matrix.

Finally, once extracted the reduced model from the cluttered sequence we need to define a *distance* among hidden Markov models in order to measure the similarity of a reduced model with one or more learned models for the action of interest. A natural way to comply is computing the *Kullback-Leibler* number between their output processes from the parameters of the models. In [9] the conjecture that the logarithms of the output sequence of HMMs have Gaussian distribution is exploited to use the integer moments of the output sequence probabilities to compute the KL number. Unfortunately, in [6] it is shown that this conjecture is wrong in the case of two simple HMMs. We decided to use a Monte-Carlo method to compute an approximated KL number, according to [5].

In conclusion we can summarize our algorithm for action detection in clutter as follows:

1. select feature components associated to the action of interest;
2. project poses onto the corresponding subspace of the feature space;
3. construct new poses by clustering;
4. create the reduced model by rearranging the topology of the graph;
5. compare the reduced model extracted from clutter with the learned one using the KL distance.

In the next section we are going to test the behavior of this technique in a simple but interesting situation and show how the results confirm our basic assumptions.

5 Experiments

As we mentioned above, the choice of a good feature representation is critical. In particular, we have to guarantee the *invariance* of our representation with respect to *translations* on the image plane and the *scaling* effect due to distance variations. Given a set of feature trajectories $\{f_1(t), ..., f_F(t)\}$, $t = 1, ..., T$ in the image plane we compute a new feature vector $y(t)$ in the following way: the bounding box for each feature trajectory is computed (Figure 1-top) and the mutual distance $r_{ij}(t)$ and orientation $\theta_{ij}(t)$ between each pair of feature at each time instant is measured (Figure 1-bottom). Calling $\{\tilde{f}_i(t)\}$ the rescaled feature coordinates with respect to the unit square and $\tilde{r}_{ij}(t)$ the mutual distances normalized by using the median of $r_{ij}(t)$ along the entire trajectory we define

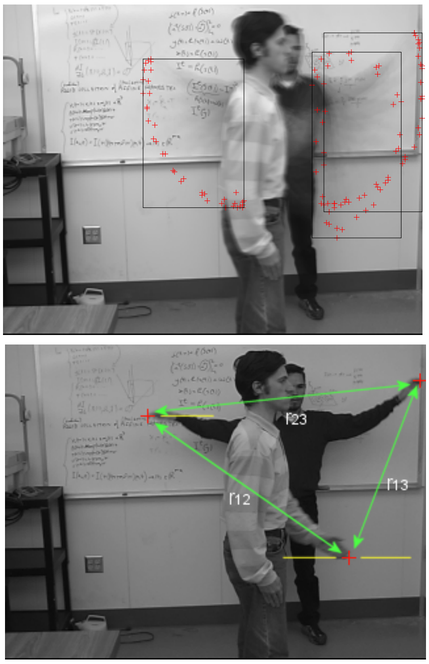

Figure 1: *Feature computation. Top: feature trajectory and bounding boxes. Bottom: mutual distances and orientations.*

$$y(t) \doteq [\{\tilde{f}_i(t)\}_{t=1}^T \ \{\tilde{r}_{ij}(t)\}_{t=1}^T \ \{\cos(\theta_{ij}(t)\}_{t=1}^T]'.$$

It can be proved that

- $\{y(t)\}_{t=1,...,T}$ is invariant with respect to translation and scaling along both axes;
- called $\mathcal{Y} : (f_1(t), f_2(t)) \mapsto y(t)$ the map transforming pairs of feature trajectories into a feature vector, the restricted application obtained by fixing one argument $\mathcal{Y}(\overline{f_1(t)}, (.))$ is injective.

In other words, fixing a scale and offset for *one* trajectory force the other into a *unique* absolute position.
To test our conjecture about the compositional properties of HMMs, we built a dataset composed by instances of three actions. "Fly", consisting on person

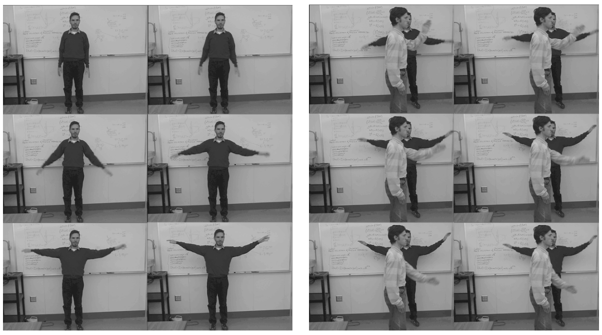

Figure 2: *Examples of motion. Left: a "fly" action. Right: combined action with "fly" and "cycle" both present and no synchronization.*

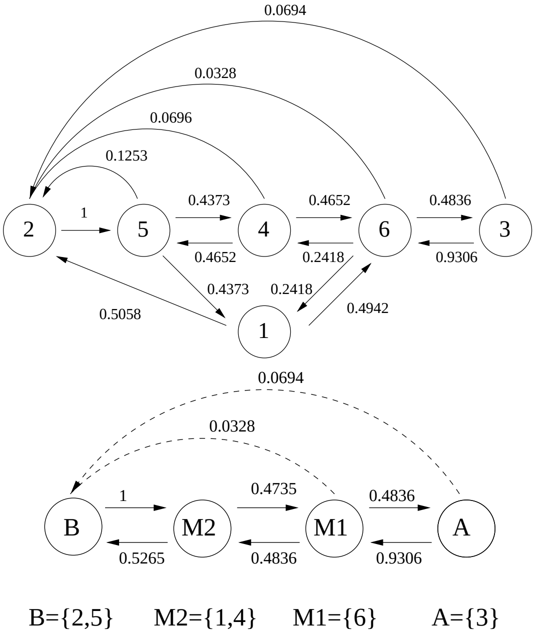

Figure 3: *Effect of the clustering on the topology of the transition matrix. Top: model for the cluttered motion in Figure 2-right. Bottom: reduced model for "fly".*

moving his arms as wings (Figure 2-left), "cycle" (a rough cycle described by a hand) and the combination of these two, executed by two people (see Figure 2-right). We implemented a simple hand tracker by means of cross-correlation filters and computed the HMM models for each of these sequences and a variable number of states: $n = 2, ..., 5$ for "fly" actions, $n = 2, ..., 4$ for "cycle" gestures and $n = 6, ..., 10$ for the instances of the combined motion. We also assumed absence of occlusions. Figure 3-top shows the graph of one of these combined models.

We applied the clustering procedure to the cluttered sequences and extracted a collection of reduced models for the "fly" gesture by selecting the appropriate feature components. Figure 3 shows the effect of the reduction algorithm on the transition matrix.

An analysis of the C matrices of both actions shows that the automatically generated clusters group together states associated to the same phase of the "fly" action. For instance, states 2 and 5 collected in the cluster B both capture the phase of "fly" in which the hands are down. The topology of the reduced model encodes almost exactly the dynamics of the action, a double chain connecting the state with hands down (B) with the state "hands up" (A) going through two intermediate positions.

The reduced models of the "fly" gestures have been calculated from the HMMs of a subset of the cluttered sequences, with number of states n=6 and n=7. Resting on our conjecture we expected both the topology and the pose matrix of these HMMs to be similar to the models learned in absence of clutter. In fact, Figure 4 shows the distribution of the poses $\{C_j,\ j = 1, ..., n\}$ of the 25 models built for "fly" with no distractors in the subspaces related to the right hand and the left hand respectively, plotted as crosses. For $n = 2$ two distinct aggregations are clearly visible, proving the stability of the model with respect to the variability of the action. On the other hand, the small squares represent the position of the poses for the reduced models for "fly" achieved by clustering from the cluttered sequences. They follow the same distribution, and the same behaviour is recognizable in the diagrams for $n = 4$. It is worth to notice that it is *not* necessary to choose a precise number of states for the cluttered model in order to extract the invariant pattern, but it suffices to have a rich enough description (i.e. $n \geq n_0$ for some n_0).

As a definitive evidence we implemented the Kullback-Leibler distance and applied it to compare the models of "cycle", "fly" and "fly in clutter" with the same number of states. The results for $n = 2$ and $n = 4$ are shown in Figure 5, and clearly confirm the similarity of reduced and a-priori models.

6 Towards unsupervised detection

The above results support our conjecture on the presence of invariant patterns of actions in clutter when actions are modeled as described in sections 2, 3. Hence we can plan the formulation of an algorithm for unsupervised detection of such models from a collection of sequences containing instances of the same *unknown* action.

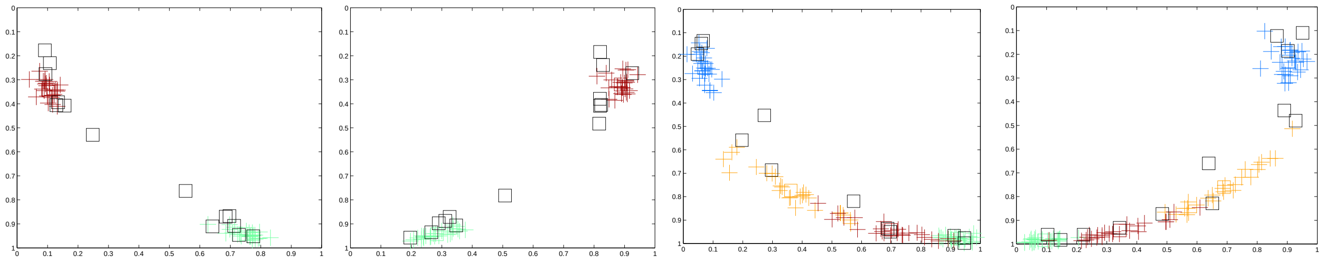

Figure 4: *Distribution of the states of "fly" (crosses) and reduced models for "fly in clutter" (squares) in the normalized feature space (unit square). From left to right: n=2, right hand; n=2, left hand; n=4, right hand; n=4, left hand.*

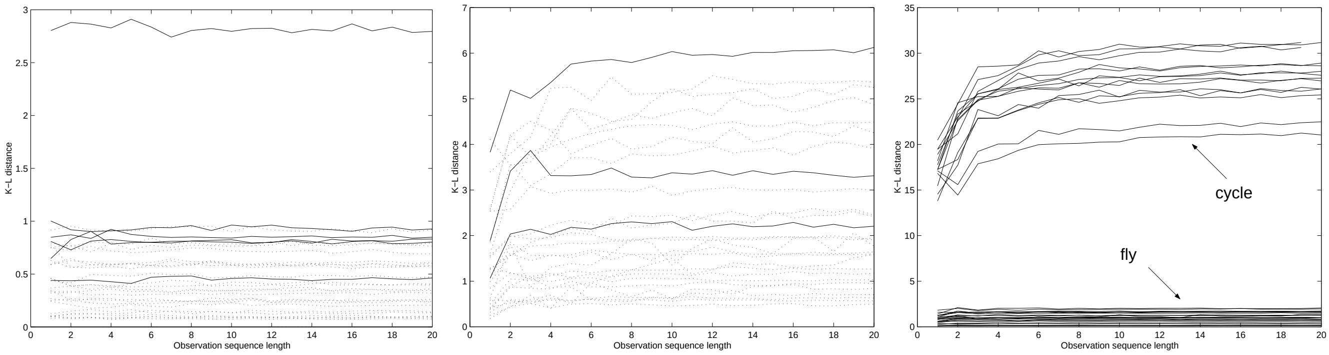

Figure 5: *Kullback-Leibler distance between models for "fly" (dotted lines) and "fly from clutter" (solid lines) and an arbitrary model for "fly" chosen as reference: the x axis plots the length of the observation sequence, the y axis the KL number. Left: n=2. Center: n=4. Right: for comparison, distance between a 3-state model for "fly" and the 3-state models for "cycle". Notice that the scales in the three plots are different and similar actions cluster closely while different actions are well separated.*

In these first tests we have assumed the absence of occlusions. Of course the problem remains critical, for standard HMM theory does not allow for observation spaces of variable dimension. A possible solution can be using standard statistical techniques for the treatment of missing data [7], based on the EM algorithm. More interesting would be the learning of models based on *hybrid* systems composed by different HMMs each representing a *state of occlusion*.

References

[1] M. Brand, N. Oliver, and A. Pentland. Coupled hmm for complex action recognition. In *Proc. of Conference on Computer Vision and Pattern Recognition*, volume 29, pages 213–244, 1997.

[2] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, pages 568–574, 1997.

[3] R. J. Elliot, L. Aggoun, and J. B. Moore. *Hidden Markov models: estimation and control.* 1995.

[4] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 22(8), pages 852–872, 2000.

[5] B. H. Juang and L. R. Rabiner. A probabilistic distance measure for hidden markov models. *AT&T Technical Journal*, Vol. 64(2):391–408, February 1985.

[6] M. Karan. *Frequency Tracking and Hidden Markov Models.* PhD thesis, 1995.

[7] J. S. Liu and Y. Wu. Parameter expansion for data augmentation. In *Journal of the American Statistical Association*, volume 94, pages 1264–1274, 1999.

[8] T. Starner and A. Pentland. Real-time american sign language recognition from video using hmm. In *Proc. of ISCV 95*, volume 29, pages 213–244, 1997.

[9] R. L. Streit. The moments of matched and mismatched hidden markov models. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 38(4):610–622, April 1990.

[10] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 21(9), pages 884–900, Sept. 1999.

[11] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. In *Computer Vision and Image Understanding*, volume 73(2), pages 232–247, 1999.