WHO IS DOING WHAT? SIMULTANEOUS RECOGNITION OF ACTIONS AND ACTORS

Muhammad Shahzad Cheema†, Abdalrahman Eweiwi†, Christian Bauckhage†‡

† B-IT, University of Bonn, Germany
‡ Fraunhofer IAIS, Germany

ABSTRACT

Recognizing human actions in videos has become a rapidly growing area of research. Most existing research has focused only on a single aspect i.e. recognition of actions. However, humans tend to perform different actions in their own styles. In this paper, we deal with the problem of simultaneously identifying actions and the underlying styles (actors) in videos. We propose a hierarchical approach based on conventional action recognition and asymmetric bilinear modeling. Our approach is solely based on dynamics of the underlying activity. Results on the multi-actor multi-action data set IXMAS show a high recognition rate.

Index Terms— Action Recognition, Bilinear Models, Motion History Volumes, Expectation Maximization

1. INTRODUCTION

People tend to execute different actions and activities such as walking, kicking and cooking in their own personal manner. It is well known from psychophysics that individuals can be recognized solely from their motion dynamics [1]. In the last decade, corresponding research on vision-based gait recognition [2] has shown great success. This motivates us to investigate the generalized multi-label classification scenario for human activity recognition. In particular, we are interested to determine if it is possible to simultaneously recognize a human action and the actor in a video? This problem is related to well-known problem of separating style from content in areas such as handwriting- or speech recognition. However, to the best of our knowledge, this is first such attempt in the context of human action recognition.

Identifying style and content has been of great interest for the recognition of handwriting, speech, and faces since the idea was pioneered in [3]. There, the authors showed promising results on classical problems such as handwritten character-, face-, or pose recognition. Related work in [4] applied a non-linear model for separating poses from walking patterns of individuals; [5] used bilinear separation models applied to different gait gestures. The approach presented in [6] was the first to consider styles of running in recognizing individuals. Despite a great deal of research on human activity recognition [7, 8, 9, 10], hardly any efforts have yet been made to separate style from content. The issue of varying styles for human activities has been addressed recently in [11] which presents a source constrained clustering approach to accommodate different sources (e.g. actors). However, the focus in [11] is on clustering from known sources and not on identifying the sources.

The approach presented in this paper treats observed actions as resulting from a generative process with two factors, namely actor (style) and action (content). We use bilinear models since they immediately lend themselves towards two-factor classification and since they can be efficiently determined through singular value decompositions (SVDs). Due to high articulation of human bodies, conventional symmetric bilinear model would not suffice to separate content and style. Notice further that the two factors, i.e. content and style, have different degrees of discrimination. It is clear from Figure 1 that discrimination among different actions (columns) is easier than among styles (rows). We therefore use a two-step approach to this multi-label problem. In the first step, we apply a classical action classifier (nearest neighbor in this paper) to predict underlying action of the query video. In the second

Fig. 1. 3D color plot in cylindrical coordinates ($r$, $\theta$, $z$) of average Motion History Volumes of 3 actions performed several times by 4 different actors in IXMAS data set.
step, we use this prediction to generate a style-specific basis for the query video using an asymmetric bilinear model. Finally, we compare this basis with a style-specific basis learned from training data in order to identify the most likely style. We show that this hierarchical model significantly improves results as compared to naive nearest neighbor approaches and symmetric bilinear models.

Next, we review the basics of bilinear models. Section 3 presents how we deal with action recognition by using nearest neighbor classifiers and asymmetric bilinear models. Details about our data set, experiments, and results are given in section 4. Finally, section 5 concludes this contribution.

2. BILINEAR MODELS

In this section, we review basic concepts of bilinear models for separating style from content; our terminology is similar to the one in [3].

A bilinear model is a generative model where each $K$ dimensional observation $y$ in a style $s \in \{1, 2, \ldots, S\}$ and content class $c \in \{1, 2, \ldots, C\}$ is given in the form:

$$y_{sc}^k = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijk} a_i^s b_j^c, \quad k \in \{1, 2, \ldots, K\}$$

(1)

where $a^s$ and $b^c$ are $I$ and $J$ dimensional coefficient vectors representing style $s$ and content $c$ and the entries $w_{ijk}$ govern interaction between the two underlying factors. The model parameters $I \leq S$ and $J \leq C$ can be chosen in different ways, e.g. by prior knowledge or for minimizing approximation error. Let $W_k$ represent $k$-th matrix of dimension $I \times J$ with entries $w_{ijk}$, then (1) can be expressed as:

$$y_{sc}^k = a^s W_k b^c$$

(2)

The matrices $W_k$ define bilinear mapping from content and style space to the $K$ dimensional observation space. The model in (1) and (2) is called the symmetric bilinear model.

While the symmetric model assumes independence of the interaction terms $w_{ijk}$ w.r.t style and content classes, the asymmetric bilinear model lets these terms vary with one of the factors (by convention with style) and thus allows for more flexibility. For instance, with a style-specific basis $a_{jk}^s = \sum_i a_i^s w_{ijk}$, equation (1) becomes:

$$y_{sc}^k = \sum_{j=1}^{J} a_{jk}^s b_j^c$$

(3)

Equivalently in matrix notation we write

$$y^{sc} = A^s b^c$$

(4)

such that $A^s$ denotes $K \times J$ matrix with entries $a_{jk}^s$. Here, $A^s$ represents a style-specific map from the content space to observation space.

2.1. Training an asymmetric bilinear model

Let $y(t)$ denote $t$-th training sample ($t = 1, \ldots, T$) and let $x_{sc}(t)$ be a characteristic function such that $x_{sc}(t) = 1$ if $y(t)$ has style $s$ and content $c$ and 0 otherwise. Then sum of squared errors $E$ for the asymmetric model over all training data is given by

$$E = \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{c=1}^{C} x_{sc}(t) \|y(t) - A^s b^c\|^2.$$  

(5)

Fitting an asymmetric model aims at finding solutions for $A^s$ and $b^c$ that minimize $E$. If a given sample of training data consists of nearly equal numbers of observations for each style and content (as in case of this paper), a closed form procedure can be adopted from using SVD.

Let $\bar{y}^s$ denotes the mean of all observations in style $s$ and content $c$, the training set can be thought of as a 3-rd order tensor $\bar{X}_{s \times K \times C}$. For making efficient use of matrix algebra, $\bar{X}$ is represented as a stacked matrix with dimensions $(SK) \times C$ such that each of $C$ columns contains $S$ chunks of $K \times 1$ vectors. Further, the vector transpose operation $VT$ is defined for stacked matrices as follows: the vector transpose of an $(AK) \times B$ matrix $Q$ is a $(BK) \times A$ matrix $Q^{VT}$ such that the $(l, m)$ entry of $Q$ becomes the $(mK + mod(l, K), l)$ entry of $Q^{VT}$.

For training data in stacked matrix form, the asymmetric model can then be expressed as $\bar{Y} = AB$, such that $A = [A^1 \ldots A^K]'$ is a $(SK) \times J$ matrix of style-specific basis and $B = [b^1 \ldots b^C]$ is a $J \times C$ matrix of content parameters. A least squares optimal solution is obtained by computing the SVD of $\bar{Y}$ such that $\bar{Y} = U \Sigma V^T$ where $\Sigma$ is diagonal matrix containing eigenvalues of $\bar{Y}$ in descending order. The style-specific basis matrix $A$ is obtained from the first $J$ columns of $US$ and the content parameter matrix $B$ is defined by the first $J$ rows of $V^T$.

2.2. Training a symmetric bilinear model

The sum of squared error for the symmetric model in (2) is

$$E = \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{c=1}^{C} x_{sc}(t) \|y(t) - A^s W_k b^c\|^2.$$  

(6)

To solve this optimization problem, asymmetric modeling through SVD is iterated by alternatively switching the roles of content and style and an expectation maximization (EM) approach is used to simultaneously update parameters of style and content. This process is based on the following relationship where the the symmetric model is given as

$$\bar{Y} = (WV^TA)^{VT}B$$

(7)

or equivalently

$$\bar{Y}^{VT} = (WB)^{VT}A$$

(8)
where $W$, $A$, and $B$ are $(IK) \times J$, $I \times S$, and $J \times C$ matrices, respectively.

An EM algorithm is used to iteratively update estimates of $A$ and $B$. The procedure starts by initializing $B$. From orthogonality of $B$ and (7), we derive $(\bar{Y}B^T)^T = W^T A$. Now, the SVD of $(\bar{Y}B^T)^T V^T = U \Sigma V^T$ is computed and the estimate for $A$ is updated to be the first $I$ rows of $V^T$. Since $A$ is orthogonal, (8) yields $(\bar{Y}^T V^T A^T)^T = W^T B$. This estimate of $A$ is used for the SVD of $(\bar{Y}^T V^T A^T)^T = U \Sigma V^T$ and $B$ is updated to be the first $J$ rows of $V^T$. This completes one iteration of the EM procedure. Upon convergence, the basis vectors are computed as

$$W = (\bar{Y}B^T V^T)^T.$$  

3. OUR APPROACH

Asymmetric bilinear models allow for more flexibility given one of the factors is known. On the other hand, symmetric models do not assume prior knowledge as to one of the factors and simultaneously update content and style parameters. Compared to asymmetric models, they are good candidates for complex tasks such as extrapolation and translation [3]. However, we observe a low performance of symmetric models for our task (see section 4).

This motivates us to use a two-stage procedure where one factor class is identified in each stage and the estimation from the first stage informs the classification in the second stage. We empirically evaluated the individual discriminativeness of the two factors by implementing single-label classification using NN approaches, i.e. by considering the problem as being either action- or actor-recognition. We observe (Fig. 2) a high accuracy for action classification (96.67%) as compared to actor-classification (47.58%).

We, therefore apply an action recognition module in the first stage of our system and use its output as the input (predicted content parameters) to the second module which is based on asymmetric bilinear models with actor-specific bases. The second module classifies the style of the query observation by using the learned basis and the predicted content class. Here, we model the training data as follows

$$\bar{Y}_{\text{train}} = A_{(SK) \times J} B_{J \times C}$$

where $A = [\tilde{A}^1 ... \tilde{A}^S]$ and $B = [b^1 ... b^C]$ are obtained as described in section 2.1.

During classification, the action-recognition module predicts an action class $\tilde{c}$ for the $K \times 1$ query observation $\tilde{y}$. In the second step, we use $\tilde{y} = \tilde{A} b^\tilde{c}$, where $b^\tilde{c}$ is $\tilde{c}$-th column of $B$ (see (10)) to determine $\tilde{A}$, i.e. the style of $\tilde{y}$. To this end, we compute

$$\tilde{A} = \tilde{y} \times (b^\tilde{c})^\dagger$$

where $(b^\tilde{c})^\dagger$ is the Moore-Penrose pseudo inverse of $b^\tilde{c}$.

Finally, we compare the $(1K) \times J$ style matrix $\tilde{A}$ for each of the $S$ chunks of $A$ and select $\tilde{s}$ such that $\arg\min_{s} |\tilde{A} - A^s|$. This procedure yields an optimal label-pair $(\tilde{s}, \tilde{c})$ for $\tilde{y}$.

4. DATA AND EXPERIMENTS

To evaluate our approach, we consider the Inria Xmas Motion Acquisition Sequences (IXMAS) [12] which is a well-known multi-actor and multi-view data set. It consists of videos of 11 actions performed 3 times by 10 different actors. Multiple executions of the same action in the same environment allow us to focus on the question if humans have unique styles for executing actions? [12] used a Fourier transform of cylindrical coordinates to express locations, scale, and rotation invariant motion history volumes from IXMAS videos as well as PCA to reduce feature space dimensionality. They report a 93% accuracy for action recognition when using Nearest Neighbor classification with the Mahalanobis distance. We used similar features with dimensionality $K = 329$. As a baseline for action-actor classification on IXMAS, we used NN classification but, instead of 11 classes, we considered 110 classes, i.e. each action-actor pair formed a separate class. With these settings we achieved around 51% accuracy. We also evaluated a symmetric bilinear model which yielded around 31% accuracy. On average, it took 60 iterations to converge to optimal style and content parameters.

In our hierarchical approach, we used action prediction to generate parameters for the asymmetric model. This yields a significantly higher accuracy of 91%. Figure 2 plots errors rates against each (action, actor) pair. It shows a non-zero error if at least one of the factors is misclassified in a given query video. Figure 3 compares the accuracies of the three
approaches across different actions. The vertical bars show style or person specific uniqueness for different actions. It shows that our approach obtains a consistent high accuracy. In particular, we observe highest style separability for the actions cross arms, getup, walk, wave, kick and pickup.

5. CONCLUSION AND FUTURE WORK

In this paper, we discussed a novel but important problem of simultaneously recognizing activities and humans in videos. We proposed a hierarchical approach based on conventional action recognition and asymmetric bilinear modeling of styles. We were able to achieve very high accuracy on a standard action recognition data set by using features invariant to location, scale and rotation. To the best of our knowledge, this is the first such investigation in the area of human action recognition. In constrained scenarios, such as surveillance or Kinect-depth imagery, our approach is directly applicable. Currently, we extend our hierarchical framework towards incorporating multiple factors including action, view, actor, scenario, camera motion, visibility, etc. This will allow us to apply our model to more realistic action data sets.

Acknowledgements

This work was carried out in the project “automatic activity recognition in large image databases” which is funded by the German Research Foundation (DFG).

6. REFERENCES