J-024

Tensor Decomposition Framework For Recognizing an Unknown Person's Action From A Video Sequence Using Image Features

Acep Irawan †　　Yingdi Xie †　　Jun Ohya †

1. Introduction

Human motion analysis is an important and challenging research topic in computer vision. This kind of gesture or action interpretation can be utilized for many applications such as video smart surveillance, cinematography and virtual reality. The aim of our research is to accurately classify the action being performed by an unknown human from his/her video sequence using a computer vision based approach, where the unknown human is not included in the database used for the classification process.

There are two research works, which are closely related to our research. First, a method proposed by Vasilescu in [1], and second proposed by our group in [3] which modified Vasilescu's method. Basically, our group's and Vasilescu's methods utilize the tensor (database), which consists of the three orders (dimensions): humans, actions and time-series motions data. The difference is that Vasilescu's method used the tensor decomposition framework using SVD (singular value decomposition) for recognizing *a known* person, who is included in the database. Since the action performed by an unknown person in the observed sequence is unknown, we assume that the observed action is one of the actions to be recognized. Then, Vasilescu's method allows us to compute the motion signature of the unknown person. All the actions of that unknown person are synthesized using the motion signature obtained from the assumption. Then, one of the persons in the original tensor is replaced by the synthesized actions, and to evaluate the appropriateness of the assumed action, the difference between the original core tensor and new core tensor is computed, where the core tensor can be obtained by applying SVD to the tensor. Here, one person's data is replaced by the synthesized actions, because if we simply append the synthesized actions to the original tensor, then the sizes of the original and new core tensors are different. After repeating this procedure for all the actions and persons, the assumption that minimizes that difference is found as the action recognition result. However, our group tested the effectiveness of this method only for time-series data obtained by a motion capture system

This paper explores the effectiveness of using image features instead of motion capture data.

2. Tensor Decomposition

Tensors, basically, are a generalization of the concept of a vector. A vector is an order-1 tensor and matrices are considered as order-2 tensors. A tensor can be considered to be a multi-dimensional or N-way array of data and as such is useful for the description of higher order quantities.

Two main steps in tensor decomposition approach are flattening or unfolding the matrix, and decompose the matrix. The main idea of a N-mode SVD derivation need to consider an appropriate generalization of the link between the column (row) vectors and the left (right) singular vectors of a matrix. To be able to formalize this idea, we define "matrix unfoldings" of a given tensor, i.e., matrix representations of that tensor in which all the column vectors are stacked one after the other.

A tensor $\mathcal{A} \in \Re^{I1 \times I2 \ldots \times IN}$ can be represented in matrix form , $\mathbf{A}_{(n)}$, which is the result of unfolding (flattening) the tensor along dimension n where $n = I_1, I_2 \ldots I_N$. Tensor unfolding can be considered as splitting a tensor into mode-n vectors and rearranging these vectors column-wise to form a matrix. In Fig. 1, a visualization is presented which demonstrates how a 3rd order tensor is unfolded along mode-1 (I_1), mode-2 (I_2) and mode-3 (I_3) dimensions to form matrices $\mathbf{A}_{(1)}$ with size $I_1 \times I_2 I_3$, $\mathbf{A}_{(2)}$ with size $I_2 \times I_3 I_1$ and $\mathbf{A}_{(3)}$ with size $I_3 \times I_1 I_2$ [2].

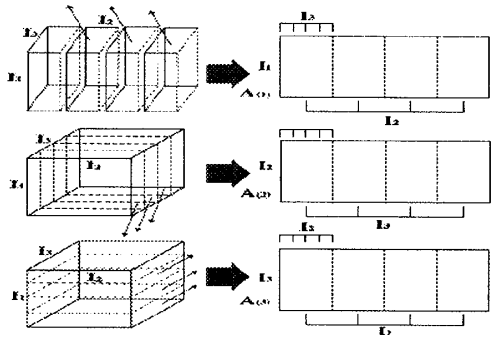

Fig. 1. The tensor can be unfolded in three ways to obtain matrices comprising of its mode-1, mode-2 or mode-3 vectors.

In addition, a tensor $\mathcal{D}$ can be expressed as a multilinear model of factor as follows [1]:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \ldots \times_n \mathbf{U}_n^T \ldots \times_N \mathbf{U}_N^T \qquad (1)$$

Where $\mathcal{Z}$, known as the core tensor, is analogous the diagonal matrix in standard SVD and U_1 to U_N contain the orthonormal vectors spanning the column space of D(n) resulting from the *mode-n flattening* of $\mathcal{D}$. Using the N-mode SVD algorithm, a multilinear extension of conventional matrix singular value decomposition (SVD), the core tensor $\mathcal{Z}$ is obtained. In case of human action data used in our experiment, the various variables are people, action and time series data. Therefore applying the SVD algorithm results in the following expression:

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{P}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{J}_3 \qquad (2)$$

where **P**, **A** and **J** denote humans, actions and time series data, respectively.

3. Images Features

This paper explores three kinds of image feature, i.e., Lt-s feature, projection feature, and mesh feature. Lt-s feature is a set of the distance between the centroid of the human silhouette

† GITS, Waseda University

and each contour pixel of the silhouette. For example, if we have a circle, the values of the data set are constant, because the distance between the centroid and the circle's contour is same. Projection feature is a set of the number of pixels in the column direction and row direction. The mesh feature is a feature vector whose element is the area ratio; i.e. the silhouette pixel's number/sub-block's pixel's number; where the image is partitioned into some smaller sub-block. We test not only single feature, but also their combinations.

To obtain the image features that should be stored in the original tensor, Lt-s curve analysis, which locates the distance in the contour of a human silhouette, is utilized. Basic concepts of Lt-s feature are as follows: video sequence is acquired, then each frame extracted . Silhouette image is obtained by subtracting the original image from background image and thresholding the subtracted image. The center of mass C is obtained by computing the centroid of the human silhouette.

To obtain the distance di between centroid and each pixel along image contour, we start obtaining P1 as start point by scanning a pixel from the centroid vertically. Let A be a contour pixel; then, Lt-s = CA + P1A. By computing the distance at each contour pixel, we obtain the Lt-s curve. Then, the Lt-s data in each frame are stored in a tensor. This process is continued until the end of all frames of video sequence.

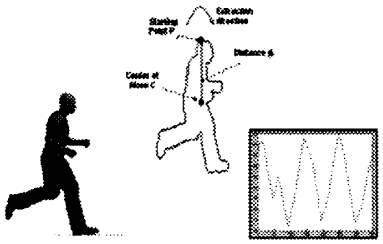

Fig. 2. A method to obtain Lt-s data. Extract video sequence, compute the distance the contour pixel and then plot the distance value.

For mesh feature; as shown in Fig. 3, suppose we have $M \times N$ pixel in the bounding box **A;** then divide A into $m \times n$ sub-blocks. The size of each sub-block is M/m by N/n pixels. On each sub block, the ratio of human silhouette pixel's number over the number of pixels in the sub-blocks is computed. Let a_{0j} (i=1,...m,j=1...n) be the ratio of the sub-block *ij*. Then, $f(a_{11},a_{12}....,a_{nm})$ is the mesh feature vector.

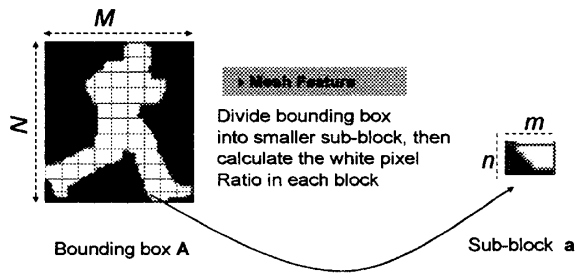

Fig. 3. Pixel ratio calculation on mesh of a silhouette image.

For projection feature, suppose we have the size of bounding box is $M \times N$ pixels. Then, as shown in Fig. 4, in each horizontal line, the number of human silhouette pixels is counted. Similarly, in each vertical line, the number of silhouette pixels is counted. Suppose Ph*i* (i=1,...N) and Pv*j* (j=1,...M) are the pixel number in i-th horizontal line and in the j-th vertical line, respectively. Then, $fp=(Ph_1,Ph_2....Ph_N, Pv_1,Pv_2...Pv_M)$ is the projection feature vector.

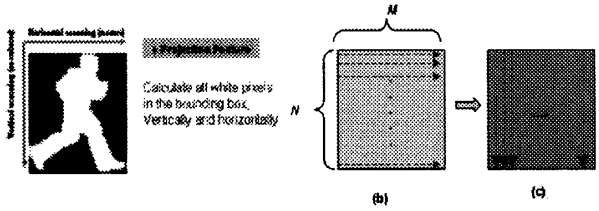

Fig. 4. Pixel calculation of projection feature.

4. Experimental Result

To evaluate our recognition method, six models are created using Curious Labs Poser ® Version 6, consist of man and woman with three kinds of actions such as normal walking, sad walking and running. We chose to have many different models, because this provides more realistic data, in addition to the fact that people have different physical characteristics, they also perform actions differently both in form and speed. We also assume that the camera is static and the one person is observed within the field of view. Each video sequence contains 25 frames. The size of each frame is 240X320 pixels.

We have tested our proposed approach not only using single image feature but also all the combinations of the feature. We also divide the database into large and small size. In the experiment we conduct the " leave -one-out " validation in which one person is chosen as an unknown person whose action is to be recognized, and the other five person's action data are used for constructing the tensor (database). Experimental results prove that the tensor decomposition framework analysis works well. We found that a single feature tends to achieve better results than combined features, while the data size of each feature turn out not to influence on the recognition accuracy significantly. Overall, the recognition accuracy is higher than 80%.

5. Summary and Conclusion

We have demonstrated an algorithm that recognizes the observed action generated by an unknown person, who is not included in the database. The results show the effectiveness of the image features used in this experiment. It turns out that the computer vision method is a significantly promising tool to recognize human action. In addition, the tensor decomposition framework has proved to be a useful method for the recognition. We also found that the data size difference of the image features does not influence on the recognition accuracy.

References:

1. M. Alex O. Vasilescu, "Human Motion Signatures : Analysis, Synthesis, Recognition", *International Conference on Pattern Recognition (ICPR'02).*
2. Lisa Grawerski, Neill Campbell,"Analysis of facial Dynamics using a Tensor Framework",Jurnal of Multimedia, Vl 1.No.5,2006.
3. Rovshan Kalanov, Jieun Cho and Jun Ohya, "A Study of Synthesizing of New Human Motions from Sampled Motions Using Tensor", ICME2005 (IEEE International Conference on Multimedia and Expo), CD-ROM Proceedings, 4 pages, (2005.7).