Head Rotation Classification Using Dense Motion Estimation and Particle Filter Tracking

Filiz Gürkan, Bilge Günsel, and Deniz Kumlu

Multimedia Signal Processing and Pattern Recognition Group, İstanbul Technical University, Turkey
{gurkan, gunselb, kumlud} @itu.edu.tr

Abstract

We propose a method that performs dense motion classification integrated with particle filter tracking for monitoring whether the viewer is involved in the screened content or not. We first perform the color based particle filtering that enables us tracking head of the user through the video sequence. It is followed by optical flow estimation via SIFT flow applied on the tracked regions. Finally the features extracted based on the viewer head rotation and location are fed into the random forest classifier to report the involvement level of the tracked person.

It is shown that the used probabilistic motion estimation model with the support of tracking significantly reduces the computational complexity while it provides comparable performance with the state-of-the-art methods. The proposed scheme allows online monitoring the viewer therefore can be integrated to the interactive multimedia systems.

1. Introduction

Head rotation estimation is used for severalapplications, such as human-computer interaction, controlling safety of driving, device control with head movement etc. In practice, the head movement estimation is a challenging task due to the difficulty to guarantee robustness to illumination changes, noise, bussy background. and occlusion. In the literature there are a number of methods proposed for head movement estimation that employ expensive sensors located in complicated measurement setups thus are mostly discomfortable for users [ref!!]. The advantage of the proposed method is we use a web camera with a simple setting.

In [1] authors present a head pose estimation system with head region detection, face detection and feature tracking by using kinect and web camera together. In this work, estimation is employs a 3D head model which represents of head shape and the relationship between the 2-D images and 3-D model. The disadvantages of this method is high computational complexity. In [2], dense Scale Invariant Feature Transform (SIFT) descriptors extracted from the detected face regions are used as representative features and the head rotation classification is performed by Support Vector Machine (SVM) classifier . Since the high dimensional SIFT vectors significantly increases the size of input data, performing supervised training of the video wiith the method presented in [2] is not feasible.

In [3] by using a single RGB camera to measure engagement level of TV viewers, first the head location is specified by Viola-Jones face detector and then facial points like eyes, nose and mouth are extracted as representative features. Obviously the proposed method relies on the performance of Viola Jones detector and in our previous work [4] we experienced that it fails especially under self-occlusion.

We propose a method for classification of the involvement level of viewer based on the head rotation. For this purpose, first the head location is determined (region of interest-ROI) via particle filter tracking [5] . This is followed by estimation of 2-D motion vectors within ROI via SIFT flow algorithm proposed in [6]. head location features, which are obtained via tracking and motion features are integrated to increase the accuracy of involvement level classification. Test results reported at Section 3 illustrate that classification integration increases the accuracy of involvement level from %66 to %72.It is also shown that the color based particle filter tracking improves robustness to occlusion.

The paper is organized as follows. Section 2 gives the theoretical background including is the SIFT flow estimation and color based particle filter tracking. Section 3 presents the proposed motion and location features that integrated to classify the head rotation. Test results are reported at Section 4 and finally, conclusions are presented in section 5.

2. Background

In this paper, main idea is integrating head location features and motion features to increase overall performance. Both head location and motion features are calculated and classified separately. Color based particle filter tracking algorithm is used for tracking head in order to obtain head location information. SIFT flow is used as a dense motion estimation algorithm.

2.1 Color Based Particle Filter Tracking

Particle filter based color trackers [5] have proved that very robust and efficient for especially non-linear and non-Gaussian estimation problems. They have superior performance for the non-rigid targets and open-world cases, whose appearance change over time. Trackers are based on deterministic search of a candidate window whose color histogram matches a predefined target histogram model. The similarity between histograms is defined by Bhattacharya distance and this metric used for updating the priori distribution calculated by the particle filter. Then, the mean of particle is found and plotted as a rectangle like in the Fig. 1.

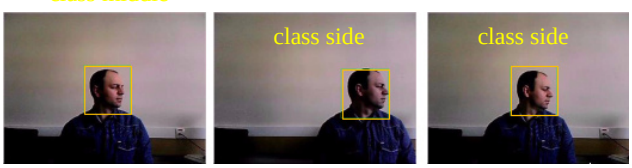

Fig. 1. Tracking results obtained by the color based particle filtering on video frames including self-occlusion. .

Viola-Jones face detector [7] is another well-known method for finding faces at each frame in the video sequence however it heavily depends on head orientation and pose of the face. In fig. 1, the head starts to rotate from center to left and color based particle filter still continues to tracking but Viola-Jones can't find the face in these frames. Color based particle filter tracking has superior performance in terms of velocity and finding location of the face compared to Viola-Jones algorithm if the pose of the head changes frequently in the video sequence. This is why we track the region of interest (ROI) by particle filtering in this work.

2.2 MAP Estimation of Motion

In this work, probabilistic optical flow method is preferred as a motion estimation method.

Image intensity can be written function of time and position: *f(x,y,t)*. According to classical gradient based optical flow total derivative of the *f(x,y,t)* must be zero.

$$(f_x f_y)^T . v + f_t = 0 \quad (1)$$

In eq 1, f_x, f_y, and f_t represent derivatives of f with respect to x,y and t and $\boldsymbol{v}$ represent flow vector.

Optical flow assumes that image intensity changes because of translation of local image intensity and lighting or noise has no effects in terms of image intensity [9]. For solution of eq 1, squared error function is written as eq 2.

$$E(v) = [(f_x f_y)^T v + f_t]^2 \quad (2)$$

To compute flow vector $\boldsymbol{v}$, we take the gradient (with respect to $\boldsymbol{v}$) and equal to 0 (eq 3).

$$\nabla E(v) = (f_x^2 \; f_x f_y \; f_x f_y \; f_y^2).v + (f_x f_t \; f_y f_t)^T = 0 \quad (3)$$

As we can see $(f_x^2 \; f_x f_y \; f_x f_y \; f_y^2)$ is a singular matrix (determinant is zero). So, it can't be solved directly. In order to eliminate this problem researchers offered different solution. But as mention before gradiant based optical flow ignores intensity changing based on lighting or noise.

Probabilistic based optical flow allows this uncertainties to be represented in computations [9]. Aim of this work is to compute an expression for the conditional probability of the image velocity based on image gradient (f_x, f_y, f_t). It is shown in eq 4.

$$P(v|(fx, fy, ft)) = \frac{P(f_t|v,(f_x f_y)).P(v)}{P(f_t)} \quad (4)$$

MAP solution of this conditional probability gives flow vector $\boldsymbol{v}$. It can be shown as eq 5.

$$v_{MAP} = argmax_v\left(P(f_t|v, f_x).P(v)\right) \quad (5)$$

For the prior distribution P($\boldsymbol{v}$), a zero-mean Gaussian with covariance λ_p is choosen [9]. To compute this probability in eq 4, eq 1 can be used with some additive gaussian noise $\boldsymbol{n_1}$ and n_2 (eq 6).

$$(f_x f_y)^T.(v - n_1) + f_t = n_2 \quad n_i \sim N(0, \lambda_i) \quad (6)$$

According to eq 5 and eq 6, energy function can be written in different ways for probabilistic based motion estimation algorithms. In our method, SIFT flow motion estimation algorithm is prefered and explained briefly in the next section.

2.3 Dense Motion Estimation by SIFT Flow

Head rotation estimation is a 3D motion estimation problem. But in this study, a method for estimating the head motion vector from a 2D face image is presented. SIFT Flow, proposed in [6], is an algorithm for estimating motion vector from video sequence. At each video frame, SIFT image is created via method which is similar to classical SIFT. Using SIFT Flow as a 2D motion estimation method, helps to improve performance of the algorithm and reduce the computational complexity.

SIFT flow is a probabilistic method which is designed similar to optical flow [8]. Unlike optical flow, SIFT flow algorithm uses SIFT image not RGB image and SIFT flow is preferred since it is more powerful and robust (Fig 2). This is the contribution of probabilistic methods which are produce useful extensions of the standard quadratic gradient techniques for computing optical flow, including an automatic gain control mechanism, and the incorporation of a prior bias on the flow. It provides (two-dimensional) flow vector confidence information, allowing later stages of processing to weight their use of the vectors accordingly, and it provides a framework for properly combining flow information with probabilistic information from other sources [9].

Fig. 2. (first row)RGB images; pan and tilt = 0° (left), pan=15° and tilt=0° (right); (second row) Reconstructed images; from SIFT flow vector (left), from Horn and Schunck optical flow vector (right)

Energy function was designed similar to that of optical flow to estimate SIFT flow. With the difference of optical flow, an additional term was added. The energy function for SIFT flow is shown in eq 7.

$$\begin{aligned} E(v) = & \sum_i \min\left(\|s_1(i) - s_2(i + v(i))\|, t\right) + \\ & \sum_i \eta(|u(i)| + |v(i)|) + \\ & \sum_{(i,j)\in\varepsilon} \min(\alpha|u(i) - u(j)|, d) + \min(\alpha|v(i) - v(j)|, d) \end{aligned} \quad (7)$$

In eq 7, *i(x,y)* is the grid coordinate of the image. In this algorithm if the grid space parameter is chosen as 1, *i* is corresponding pixels of the image. In equation 1, *v(i)=[u(i),v(i)]* is a flow vector in *p*. Besides s_1 and s_2 represent SIFT images

which are nxmx128 dimension for nxmx3 RGB image. *t* and *d* are threshold while η and α are coefficients of equation.

First term of Eq 7, 'data term', constrains the SIFT descriptors to be matched with the flow vector Second term, 'small displacement term' constrains the flow vectors to be as small as possible. Third term, 'smoothness term' constrains the flow vectors in of neighbor pixels to be similar [6]. ε is neighborhood parameter and chosen as 4 (four neighbor system) in this work.

The optimum solution of energy function in Eq 7 is called flow vectors '***v***'. Belief propagation algorithm is used to optimize and get flow vector from energy function[6].

3. Integration of Location and Motion Features

SIFT Flow motion estimation, with the difference of optical flow, is using SIFT images (fig 3) to estimate flow vector. Each pixel is represented with 128-D SIFT descriptor in SIFT image [6]. Using SIFT image for estimation gives robustness against to lighting and noise without increasing complexity and run-time of the algorithm.

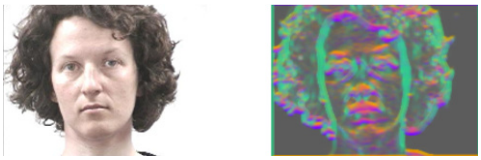

Fig. 3. (left) RGB image; (right) SIFT image (128D image reduce 3D by using PCA to representation)

The color based particle filter tracking enable us specifying the region of interest, which is head of the user for our experiment, in the monitoring video sequence. This intercepts false motion vector cause by camera motion and head movement. Also, calculating flow vector in the background is not giving any valuable information for the head rotation and increases the run-time of algorithm.

Motion flow vectors are calculated in specified region via color based particle filter tracking algorithm. In order to calculate motion vectors in video sequence, a target frame should be assigned as a reference frame which is the first frame of video sequence in this work. All flow vectors and motion features are calculated between target frame and the other frames individually. Head motion feature vectors which are amplitude and angle of flow vectors ($v(i)=[u(i),v(i)]$) extracted as shown in eq 8.

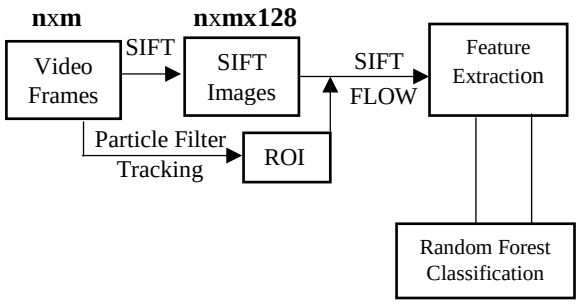

Fig. 4. Feature extraction.

$$amp(i) = \sqrt{u(i)^2+v(i)^2} \qquad ang(i) = tan^{-1}\frac{u(i)}{v(i)} \qquad (8)$$

For *k x z* dimension region of interest, there are *k x z* observations with 2D feature vector. In order to avoid more run-time and complexity, which is caused by more observations, grid space parameter is changed from 1 to 2. Thus, one flow vector is calculated for 4 pixel group. So both observation number and run-time of the algorithm decreases at the rate of 0.25 without reducing the performance.

For head location feature, 3-D features (*x, y,A)*, are used. *(x,y)* is the beginning coordinates of the location and *A* is the area of the head location which is found via color based particle filter tracking algorithm (fig 5).

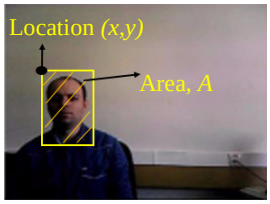

Fig. 5. Head location features

4. Test Results

The aim of this work is integrating head location features and motion features for monitoring whether the viewer is involved in screen content or not. Before calculating the motion vectors, ROI is specified by color based particle filter tracking algorithm and head location features are used to classify head location as in the 'middle' or 'side. After that, motion estimation is applied on SIFT image in order to calculate motion vectors. It is assumed that, frames which are classified as a 'side', are 'not involved' frames. So motion estimation is only applied to frame which is labelled as 'middle'. Since, the main goal of this work is determining involved frames

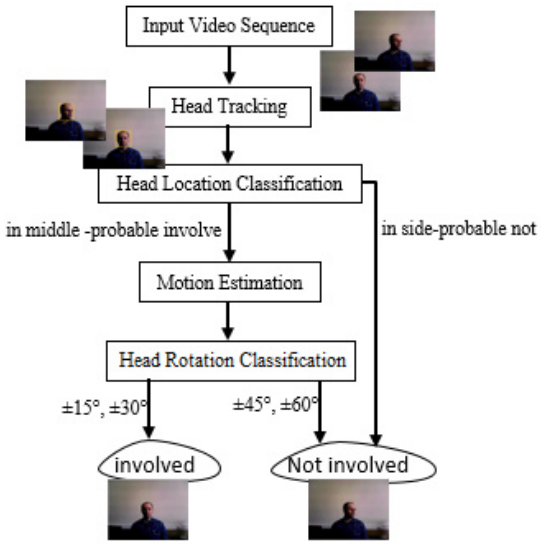

Fig. 6. Head motion and head location integration algorithm for head rotation classification

The motion classification problem is a binary classification problem. Therefore, we evaluated 28 binary datasets from 8 classes of the pan movement (x direction). To classify, each video frame classifies with each binary dataset. After, classification part is finalized, video frames are labeled in terms of angle class with using majority voting algorithm.

According to angle class labels, it is decided involvement level as “involve” or “not involve”. Class *±15* or *±30* degree accepted as “involve”, whereas class *±45* or *±60* degree accepted as “not involve” for our experiment. This algorithm is shown fig 6.

Random Forest classification algorithm [10] is preferred as a classifier algorithm since it has low computational complexity and high speed. Random forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution of all trees in the forest. In this case 10 was chosen as a number of tree. Classification part of this study was made in WEKA [11].

4.1 Dataset and Test Cases

In motion estimation problem, two motion classes are proposed. Class ‘involved’ comprise *±15°*,*±30°* head rotation angles in pan direction and class ‘not involved’ comprise *±45°*,*±60°* head rotation angles in pan direction (Fig 7).

Fig. 7. Head rotations corresponding to. +60°, -45° (not involved), and +30° (involved).

As a training set in motion classification part, Pointing’04 database [12] is shown in fig 7. In this database there are 15 different people and for each of them there are 9 different angles (*0°*,*±15°*,*±30°*,*±45°*,*±60°*). In other words train dataset involves 120 frames and each frame has a number of features as much as their pixels. So, in training set there are about 214331 features. 142771 of these features were labelled as an ‘involved’ and 71560 of them were labelled as a ‘not involved’.

In head location classification, there are two classes which are ‘middle’ and ‘side’. In involvement level classification problem, class of middle corresponds to ‘probably involved’ and class of side corresponds to ‘probably not involve’. Middle and side classes’ criteria are shown in Table 1. This criterion is constituting just for frames which is 480x640 dimension.

Table 1. Head location class descriptions for an (480x640) image((*x*,*y*) refers to an image pixel).

Middle (probably involved)	Side (probably not involved))
149< x < 351 50< y < 231	x < 150 or x > 350 y < 51 or y > 230

As a training set in head location classification, real videos which have 12460 frames were used (Fig 8). In this training set 6017 frames are labeled as ‘middle’ and other frames are labeled as ‘side’.

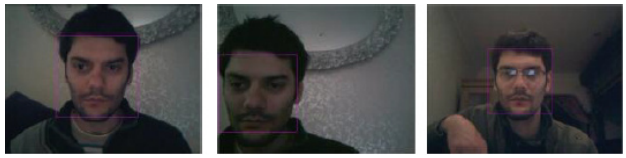

Fig. 8. Representative video frames used for the location based training.

For the test set, real video is recorded with dimension *480x640* (Fig 9). This video has 270 frames with different head angles and locations. In order to measure the performance in that video, video is labeled by manually as an ‘involved’ or ‘not involved’.

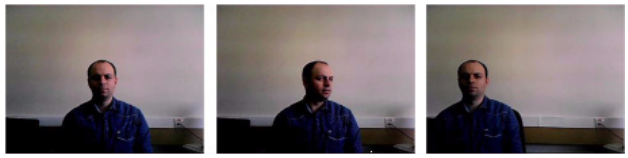

Fig. 9. Test set video frames

4.2 Numerical Results

The measure of the performance used in recall (R), precision(P) and F measure (F), which is the harmonic mean between precision and recall. Precision is defined as the fraction of retrieved instances that are relevant while recall is defined the fraction of relevant instances that are retrieved. These parameter’s calculations are shown in below.

$$P_i = \frac{TC_i}{TN_i} \quad R_i = \frac{TC_i}{CN_i}$$

$$F_i = 2\,x\,\frac{P_i\,xR_i}{P_i + R_i} \qquad i = \{involved,not\ involved\} \qquad (9)$$

In these equations, TC_i is the number of true classifying in class i, TN_i is the total number of observations in class i and CN_i is the number of observation which is classified in class i.

In this work, it is expected that the proposed method, which is about integrating motion classification and location classification, increases the classification performance. In order to prove that claim, classification performance was measured with two different tests. In one of the test, which is called ‘test 1’, only SIFT flow is performed for all video frames. In the other test, which is called ‘test 2’, motion classification and location classification is integrated.

In test 1, the color based particle filter tracking is used only to specify the location of head as a ROI. This means that all the frames classified according to motion features as ‘involved’ or ‘not involved’.

In test 2, the color based particle filter tracking is used both to specify the ROI and to classifiying head location as ‘middle-probably involved’ or ‘side-probably not involved’.

For test 1, SIFT flow and tracking algorithms are performed [13] and then motion features are calculcated for all 270 frames' ROI. Each frame's features are classified seperatly. Classification results are shown in table 1 as a test 1. For test 2, color based particle filter algorithm is performed and head location features are calculcated for all 270 frames. Video frames, which are classified 'middle' in location classification, are labeled 'probably involve' and frames, which are classified 'side', are labeled 'probably not involve'. For our test video sequence, 199 frames were labeled as 'middle-probably involved'. In our approach, it can be said that frames, which were labeled as 'side', can be labeled as 'not involved' without motion classification.

Thus, SIFT flow algorithm is run only for 199 frames, which are labeled as 'middle' via head location classification. After motion features are obtained, motion classification is performed with random forest classifier for 199 frames. Classification results are shown in table 1 as a test 2.

As mentioned earlier one of the main advantages of the integration is that run-time of the algorithm considerably decreases. Since some of the frames are labeled 'not involved' via location estimation and motion features aren't calculated for these frames. It is clear that performing motion estimation just for 'middle' frames, helps to decrease test time by approximately %30.

In both test 1 and test 2, first frame of video is defined as a target frame.

Recall, precision and F measure parameters are used to report the classification performances.

Table 2. Test results for only motion features and integration of motion and head location features according to involved class performance

	Test 1	Test 2
Recall	0,68	0,82
Precision	0,64	0,64
F measure	0,66	0,72

F measure result for test 1 is %66 which is similar to results in literature works. For test 2, performance is increased to %72, while run-time is decreased as we expected.

When the proposed approach is compared with method presented in [4], test 2 result is better than [4]. Also, in [4], Viola-Jones face detector was used for head location. As we have mentioned before, Viola-Jones algorithm can't be found ROI in all frames unlike particle filter tracking.

5. Conclusions

In this paper, we have presented a method that monitors involvement level of the viewer with a single camera. By using SIFT image instead of RGB image in motion estimation, we provide robustness against to noise and illumintion changes.

The motion vectors are calculated using SIFT flow algorithm in a specified head location which is tracked by the color based particle filter tracking. SIFT flow is the probabilistic motion estimation method which gives more reliable results than gradient optical flow algorithm.

The viewer head motion and location features are classified by random forest classifier. Experimental results demonstrate the effectiveness of the proposedmethod to achieve involvement level monitoring by using a single RGB camera with a small cost. Furthermore, with the help of reliable 2D motion estimation by SIFT flow we significantly reduce the computational complexity that enable us near real time monitoring. Currently we are working on non-uniform tracking to minimize the computational load.

Although our goal is measuring involvement level of the viewer to the screened content, similar systems can be used for many other applications including driver safety monitoring, interactive computer games, etc.

6. References

[1] P. Liu, M. Reale, and L. Yin. "3D Head Pose Estimation Based on Scene Flow and Generic Head Model", *IEEE Int. Conf. Multimed. Expo*, pp. 794–799, 2012.

[2] H.T. Ho and R. Chellappa, "Automatic Head Pose Estimation Using Randomly Projected Dense SIFT Descriptors", *IEEE Image Processing (ICIP)*, pp. 153-156 , 2012

[3] J. Harnandez, Z. Liu, G. Hulten, D.DeBarr, K. Krum and Z. Zhang. "Measuring the Engagement Level of Tv Viewers", *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1-7 , 2013

[4] F. Gürkan, B. Günsel, and E. Başyurt, ""Integration of Motion and Localization Features for Head Movement Classification", , *Signal Processing and Communications Applications Conference* pp. 1586- 1589, 2015

[5] Nummiaro, Katja, Esther Koller-Meier, and Luc Van Gool. "An adaptive color-based particle filter." Image and vision computing", 21(1), 99-110, 2003.

[6] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "SIFT Flow : Dense Correspondence across Different Scenes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 1, pp. 28–42, 2008.

[7] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[8] B. K. P. Horn and B. G. Schunck, "'Determining optical flow': a retrospective," *Artif. Intell.*, vol. 59, no. 1–2, pp. 81–87, 1993.

[9] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability Distributions of Optical Flow", IEEE Conference on Computer Vision and Pattern Recognation, pp. 310-315, 1991.

[10] L. Breiman, "Random Forrest," Mach. Learn., pp. 133, 2001.

[11] url 2 < http://www.cs.waikato.ac.nz/ml/weka/>

[12] url 1 < http://www-prima.inrialpes.fr/Pointing04 /data-face.html>

[13] url 3 <http://people.csail.mit.edu/celiu/SIFTflow/>