Egocentric Multimedia for healthcare applications

Pr. Jenny Benois-Pineau,
LaBRI – Université de Bordeaux – CNRS UMR 5800/ University Bordeaux/IPB-ENSEIRB
H. Boujut, V. Buso, L. Letoupin, R. Megret, S. Karaman, V. Dovgalecs, (UBx), I. Gonzalez-Diaz (UHCM)
Y. Gaestel, J.-F. Dartigues (INSERM)
Summary

1. Introduction and motivation
2. Egocentric video
3. Fusion of multiple cues in H-HMM for IADL Recognition
4. Object recognition with saliency Maps from egocentric video.
5. Conclusion and perspectives
1. Introduction and motivation

- Egocentric vision since mainly 2000th
- Objective assessment of capacities on instrumental activities of daily living of persons with Dementia
- Thus recognition of activities and Manipulated objects in EC video
- Multidisciplinary projects: IT/Multimedia Research and Medical Research
  ANR IMMED https://immed.labri.fr/
  IP FP7 EU Dem@care http://www.demcare.eu/
2. Egocentric Video

- Video acquisition setup

- Wide angle camera on shoulder
- Non intrusive and easy to use device
- IADL capture: from 40 minutes up to 2.5 hours
Egocentric Video

• 4 examples of activities recorded with this camera:
• Making the bed, Washing dishes, Sweeping, Hovering
3. Fusion of multiple cues in H-HMM for IADL Recognition

3.1 General architecture
3.2. Temporal Partitioning (1)

- Pre-processing: preliminary step towards activities recognition
- Objectives:
  - Reduce the gap between the amount of data (frames) and the target number of detections (activities)
  - Associate one observation to one viewpoint
- Principle:
  - Use the global motion e.g. ego motion to segment the video in terms of viewpoints
  - One key-frame per segment: temporal center
  - Rough indexes for navigation throughout this long sequence shot
  - Automatic video summary of each new video footage
Temporal Partitioning(2)

- Complete affine model of global motion (\(a_1, a_2, a_3, a_4, a_5, a_6\))

\[
\begin{pmatrix}
\frac{dx_i}{dt} \\
\frac{dy_i}{dt}
\end{pmatrix} =
\begin{pmatrix}
a_1 \\
a_4
\end{pmatrix} +
\begin{pmatrix}
a_2 & a_3 \\
a_5 & a_6
\end{pmatrix}
\begin{pmatrix}
x_i \\
y_i
\end{pmatrix}
\]


- Principle:
  - Trajectories of corners from global motion model
  - End of segment when at least 3 corners trajectories have reached outbound positions
Temporal Partitioning (3)

- Threshold $t$ defined as a percentage $p$ of image width $w$

  $p = 0.2 \ldots 0.25$

  $$t = p \times w$$
Temporal Partitioning(4)

Video Summary

• 332 key-frames, 17772 frames initially
• Video summary (6 fps)
3.3 Description space: fusion of features (1)

- Color: MPEG-7 Color Layout Descriptor (CLD)
  6 coefficients for luminance, 3 for each chrominance
  - For a segment: CLD of the key-frame, $x(\text{CLD}) \in \mathbb{R}^{12}$

- Localization: feature vector adaptable to individual home environment.

- $N_{\text{home}}$ localizations. $x(\text{Loc}) \in \mathbb{R}^{N_{\text{home}}}$

- Localization estimated for each frame

- For a segment: mean vector over the frames within the segment

- Audio: $x(\text{Audio})$ : probabilistic features SMN…

Description space(2). Audio

- **Silence detection**
  - Energy
  - Silence probability

- **Speech detection**
  - Entropy modulation
  - 4 Hz energy modulation
  - Fusion (scores)
  - Speech probability

- **Music detection**
  - Number of segments
  - Segment duration
  - Fusion (scores)
  - Music probability

- **Noise detection**
  - MFCC with GMM
  - Other noise probabilities
  - Spectral cover
  - Water flow and vacuum cleaner probabilities

7 audio confidence indicators

J. Pinquier, IRIT
Description space(3). Motion

- $H_{tpe}$ log-scale histogram of the translation parameters energy
  Characterizes the global motion strength and aims to distinguish activities with strong or low motion

- $N_e = 5$, $s_h = 0.2$. Feature vectors $x(H_{tpe}, a_1)$ and $x(H_{tpe}, a_4) \in \mathbb{R}^5$
  \[
  H_{tpe}[i] = 1 \quad \text{if} \quad \log(a^2) < i \times s_h \quad \text{for} \quad i = 1
  
  H_{tpe}[i] = 1 \quad \text{if} \quad (i - 1) \times s_h \leq \log(a^2) < i \times s_h \quad \text{for} \quad i = 2..N_e - 1
  
  H_{tpe}[i] = 1 \quad \text{if} \quad \log(a^2) \geq i \times s_h \quad \text{for} \quad i = N_e
  
- Histograms are averaged over all frames within the segment

<table>
<thead>
<tr>
<th></th>
<th>$x(H_{tpe}, a_1)$</th>
<th>$x(H_{tpe}, a_4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low motion segment</td>
<td>0.87 0.03 0.02 0 0.08</td>
<td>0.93 0.01 0.01 0 0.05</td>
</tr>
<tr>
<td>Strong motion segment</td>
<td>0.05 0 0.01 0.11 0.83</td>
<td>0 0 0 0.06 0.94</td>
</tr>
</tbody>
</table>
Description space(4). Motion

• $H_c$: cut histogram. The $i^{th}$ bin of the histogram contains the number of temporal segmentation cuts in the $2^i$ last frames


• Average histogram over all frames within the segment

• Characterizes the motion history, the strength of motion even outside the current segment

$2^6=64$ frames $\rightarrow$ 2s, $2^8=256$ frames $\rightarrow$ 8.5s $x(H_c) \in \mathbb{R}^6$ or $\mathbb{R}^8$

• Residual motion $\quad RM_b = \sqrt{\frac{\sum_{k=1,l=1}^{N,M} (\Delta x_{k,l}^2 + \Delta y_{k,l}^2)}{N \times M}} \quad 4 \times 4 \times x(RM) \in \mathbb{R}^{16}$
Description space(5). Place recognition

Granularity of visual recognition

**Room recognition** [Dovgalecs 2010]

Topological positioning [O’Conaire 2009]

**Metric 3D positioning** [Wannous 2012]
Evaluation of Room Recognition

On the IMMED dataset @Home

Difficult dataset due to the low amount of training data for each location (5 minutes bootstrap)

Conclusions

Best performances obtained using late-fusion approaches with temporal accumulation

Average accuracy for room recognition on IMMED dataset.

BOVW - Wearable camera video analysis
Early fusion of all features

- Dynamic
- Static
- Audio

Combination

HMM Classifier

Activities
### Description space (6)

- **#Possible combinations of descriptors**: \(2^6 - 1 = 63\)

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Audio</th>
<th>Loc</th>
<th>RM</th>
<th>Htpe</th>
<th>Hc</th>
<th>CLD</th>
<th>config min</th>
<th>config max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensions</td>
<td>7</td>
<td>7</td>
<td>16</td>
<td>10</td>
<td>8</td>
<td>12</td>
<td>7</td>
<td>60</td>
</tr>
</tbody>
</table>
3.4 Model of the content

HMMs: efficient for classification with temporal causality
An activity is complex, it can hardly be modeled by one single state
Hierarchical HMM? [Fine98], [Bui04]

- Multiple levels
- Computational cost/Learning

• $Q^D = \{q_i^d\}$ states set

$$\Pi_{q_i^d} (q_{j^{d+1}}) = \text{initial probability of child } q_{j}^{d+1} \text{ of state } q_i^d$$

• $A_{ij}^q = \text{transition probabilities between children of } q^d$
Model of the content: activities recognition

A two level hierarchical HMM:

- Higher level:
  - Example activities: Washing the dishes, Hovering, Making coffee, Making tea...

- Bottom level:
  - Activity: HMM with 3/5/7/8 states
  - Observations model: GMM
  - Prior probability of activity
Activities recognition

Bottom level HMM

- Start/End
  → Non emitting state
- Observation x only for emitting states $q_i$
- Transitions probabilities and GMM parameters are learnt by Baum-Welsh algorithm
- A priori fixed number of states
- HMM initialization:
  - Strong loop probability $a_{ii}$
  - Weak out probability $a_{i\text{end}}$
### 3.5. Results

#### Video Corpus

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Healthy volunteers/Patients</th>
<th>Number of videos</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMMED</td>
<td>12 healthy volunteers</td>
<td>15</td>
<td>7H16</td>
</tr>
<tr>
<td>IMMED</td>
<td>42 patients</td>
<td>46</td>
<td>17H04</td>
</tr>
<tr>
<td>TOTAL</td>
<td>12 healthy volunteers + 42 patients</td>
<td>61</td>
<td>24H20</td>
</tr>
</tbody>
</table>
Evaluation protocol

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{precision} = \frac{TP}{TP+FP}$</td>
<td>$\text{recall} = \frac{TP}{TP+FN}$</td>
</tr>
<tr>
<td>Accuracy</td>
<td>F-score</td>
</tr>
<tr>
<td>$\text{accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)}$</td>
<td>$\text{F-score} = \frac{2}{\frac{1}{\text{precision}}+\frac{1}{\text{recall}}}$</td>
</tr>
</tbody>
</table>

- Leave-one-out cross validation scheme (one video left)
- Results are averaged
- Training is performed over a sub sampling of smoothed (10 frames) data.
- Label of a segment is derived by majority vote of frames results
Recognition of Activities
5 videos. Descriptors?

3 states LL HMM,
1 state None
Comparison with a GMM baseline (23 activities on 26 videos)
« Hovering - specific audio description »
Conclusion on early fusion

Early fusion requires tests of numerous combinations of features.

The best results were achieved for the complete description space.

For specific activities optimal combinations of descriptors vary and correspond to « common sense » approach.

Intermediate fusion

Treat the different modalities (Dynamic, Static, Audio) separately.

We represent each modality by a stream, that is a set of measures along the time.

Each state of the HMM models the observations of each stream separately by a Gaussian mixture.

\( K \) streams of observations \( O_{i,1}, \ldots, O_{i,k} \) \( O_{i,k} \in \mathbb{R}^{N_k} \)

\[ \sum N_k = N \]

\[ p(o_i, q_j) = \prod_{k=1}^{K} p_k(o_{i,k}, q_j)^{w_{lk}} \]
Late Fusion

MEDIA

Dynamic Features

Static Features

Audio Features

HMM Classifier

HMM Classifier

HMM Classifier

Fusion of scores of classifiers

Activities

Performance measure of classifiers:
modality $k$, activity $l$

$$e_{lk} = \frac{perf_{lk}}{\sum_k perf_{lk}}$$
Experimental video corpus in 3-fusion experiment

37 videos recorded by 34 persons (healthy volunteers and patients) for a total of 14 hours of content.
Results of 3- fusion experiment (1)
Results of 3- fusion strategies(2)

<table>
<thead>
<tr>
<th>Metrics (averaged)</th>
<th>Early fusion</th>
<th>Interim. fusion</th>
<th>Late Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>F-score trust</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.207</td>
<td>0.442</td>
<td>0.215</td>
</tr>
<tr>
<td>Precision</td>
<td>0.174</td>
<td>0.267</td>
<td>0.106</td>
</tr>
<tr>
<td>Recall</td>
<td>0.284</td>
<td>0.288</td>
<td>0.171</td>
</tr>
<tr>
<td>F-Score</td>
<td>0.180</td>
<td>0.253</td>
<td>0.109</td>
</tr>
</tbody>
</table>
Conclusion on 3-fusion experiment

Overall, the experiments have shown that the intermediate fusion has provided consistently better results than the other fusion approaches, on such complex data, supporting its use and expansion in future work.
4. Object recognition with saliency Maps from wearable video.

- Why?
- Saliency Modeling
- Object recognition in egocentric videos with saliency
- Results
- Conclusion
INTRODUCTION

- Object recognition
- From wearable camera
- Egocentric viewpoint
- Manipulated objects from activities of daily living
OBJECT RECOGNITION WITH SALIENCY

• Many objects may be present in the camera field
• How to consider the object of interest – “active object”? 
• Our proposal: By using visual saliency. This is a popular subject!
VISUAL ATTENTION

• Several approaches
  • Bottom-up or top-down
  • Overt or covert attention
  • Spatial or spatio-temporal
  • Scanpath or pixel-based saliency

• Features
  • Intensity, color, and orientation (Feature Integration Theory [1]), HSI or L*a*b* color space
  • Relative motion [2]

• Plenty of models in the literature
  • In their 2012 survey, A. Borji and L. Itti [3] have taken the inventory of 48 significant visual attention methods

Saliency Model

SOA: spatial and temporal, contribution «egocentric» geometry fusion: least square optimisation wrt OR score:

\[ S(p) = \alpha Sp(p) + \beta St(p) + \gamma Sg(p), \alpha + \beta + \gamma = 1 \]

2. VBuso, J Benois-Pineau, I Gonzalez-Diaz, « Object recognition in egocentric videos with saliency-based non uniform sampling and variable resolution space for features selection, CVPR’2014, 3rd WS on Egocentric (First-person) Vision."
The saliency peak is never located on the visible part of the shoulder.

Most of the saliency peaks are located on the 2/3 at the top of the frame.

So the 2D Gaussian center is set at:

\[ x_0 = \frac{\text{width}}{2} \quad y_0 = \frac{\text{height}}{3} \]
GEOMETRIC SALIENCY MODEL

- 2D Gaussian was already applied in the literature [1]
  - “Center bias”, Busswel, 1935 [2]
  - Suitable for edited videos

- Our proposal:
  - Train the center position as a function of camera position
  - Move the 2D Gaussian center according to camera center motion.

Geometric saliency map
Saliency Fusion

Frame
Spatio-temporal-geometric saliency map
Subjective saliency map
PROPOSED PROCESSING PIPELINE FOR OBJECT RECOGNITION

Spatially constrained approach using saliency methods
Contribution of specific egocentric cues, CHU Nice dataset

CHU Nice dataset 44 videos, 9h 30 min, 17 categories of active objects, object occurrences 102 (tea box) – 2032 (tablet), 22236 frames annotated
CONCLUSION

• Activities recognition with fusion of low-level and mid-level features
• Object recognition with saliency models specifically adapted to egocentric video
• Work in progress: new model of activity = objects + location
• Fusion of egocentric view and 3rd person view for activities recognition
**Spatial Saliency Model**

- Based on the sum of 7 color contrast descriptors in HSI domain [1][2]
  - Saturation contrast
  - Intensity contrast
  - Hue contrast
  - Opposite color contrast
  - Warm and cold color contrast
  - Dominance of warm colors
  - Dominance of brightness and hue

The 7 descriptors $V_\delta$ are computed for each pixel $s_i$ of a frame $I$ using the 8 connected neighborhood.

The spatial saliency map $S^{SP}$ is computed by:

$$S^{SP}(s_i) = \frac{1}{7} \sum_{\delta=1}^{7} V_\delta(s_i)$$

Finally, $S^{SP}$ is normalized between 0 and 1 according to its maximum value $S_{max}$

$$S^{SP'}(s_i) = \frac{S^{SP}(s_i)}{S_{max}}$$


The temporal saliency map is extracted in 4 steps [Daly 98][Brouard et al. 09][Marat et al. 09]

1. The optical flow is computed for each pixel $s_i$ of frame $i$.
2. The motion is accumulated in $\overrightarrow{V}_0(s_i)$ and the global motion $\overrightarrow{V}_G(s_i)$ is estimated.
3. The residual motion is computed:
   \[
   \overrightarrow{V}_R(s_i) = \overrightarrow{V}_0(s_i) - \overrightarrow{V}_G(s_i)
   \]
4. Finally, the temporal saliency map $S^T(s_i)$ is computed by filtering the amount of residual motion in the frame.

\[
S^T(s_i) = \begin{cases} 
\frac{1}{7} \overrightarrow{V}_R(s_i) & \text{if } 0 \leq \overrightarrow{V}_R(s_i) < \tilde{v}_1 \\
1 & \text{if } \tilde{v}_1 \leq \overrightarrow{V}_R(s_i) < \tilde{v}_2 \\
\frac{1}{60} \overrightarrow{V}_R(s_i) + \frac{9}{5} & \text{if } \tilde{v}_2 \leq \overrightarrow{V}_R(s_i) < \tilde{v}_{\text{max}} \\
0 & \text{if } \overrightarrow{V}_R(s_i) \geq \tilde{v}_{\text{max}}
\end{cases}
\]

with $\tilde{v}_1 = 6\text{ deg./s}$, $\tilde{v}_2 = 30\text{ deg./s}$ and $\tilde{v}_{\text{max}} = 80\text{ deg./s}$