Tri-modal Human Body Segmentation

Master of Science Thesis

Cristina Palmero Cantariño

Advisor: Sergio Escalera Guerrero

HuPBA

February 6, 2014

Outline

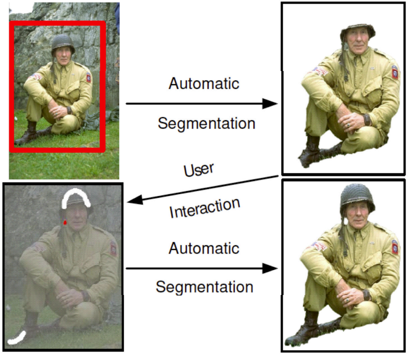
Outline

Human body segmentation

Segmentation: labeling problem.

Main challenges:

- Different points of view.
- Illumination changes.
- Complex and cluttered backgrounds.
- Presence of occlusions.
- Human body articulated nature.
- Diversity of poses.
- Variable appearance.



Segmentation using Grabcut

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "Grabcut: Interactive foreground extraction using iterated graph cuts". In: *ACM Transactions on Graphics (TOG)*. Vol. 23. 3. ACM. 2004, pp. 309–314.

Motivation

Applications:

- Security.
- Leisure.
- Health.

Imaging modalities:

- Mostly RGB cues from color cameras.
- Recently, RGB-Depth cues (Microsoft® Kinect™).
- Little attention to thermal.

Thermal Imaging:

- Price of thermal sensors is lowering substantially every year.
- Less intrinsic problems than RGB cues.
- Lack of benchmarks comparing RGB-Depth-Thermal modalities.

Motivation

Applications:

- Security.
- Leisure.
- Health.

Imaging modalities:

- Mostly RGB cues from color cameras.
- Recently, RGB-Depth cues (Microsoft®Kinect™).
- Little attention to thermal.

Thermal Imaging:

- Price of thermal sensors is lowering substantially every year.
- Less intrinsic problems than RGB cues.
- Lack of benchmarks comparing RGB-Depth-Thermal modalities.

Motivation

Applications:

- Security.
- Leisure.
- Health.

Imaging modalities:

- Mostly RGB cues from color cameras.
- Recently, RGB-Depth cues (Microsoft®Kinect™).
- Little attention to thermal.

Thermal Imaging:

- Price of thermal sensors is lowering substantially every year.
- Less intrinsic problems than RGB cues.
- Lack of benchmarks comparing RGB-Depth-Thermal modalities.

Proposal

Tri-modal dataset

- Novel tri-modal dataset of continuous image sequences.
- RGB-Depth-Thermal modalities.
- People interacting with everyday objects.

Baseline methodology

- Automatic segmentation of people in video sequences in indoor scenarios with a fixed camera.
- Usage of state-of-the-art descriptors for feature extraction among modalities.
- GMM modeling of subject/object regions.
- Multi-modal fusion using several approaches.

Proposal

Tri-modal dataset

- Novel tri-modal dataset of continuous image sequences.
- RGB-Depth-Thermal modalities.
- People interacting with everyday objects.

Baseline methodology

- Automatic segmentation of people in video sequences in indoor scenarios with a fixed camera.
- Usage of state-of-the-art descriptors for feature extraction among modalities.
- GMM modeling of subject/object regions.
- Multi-modal fusion using several approaches.

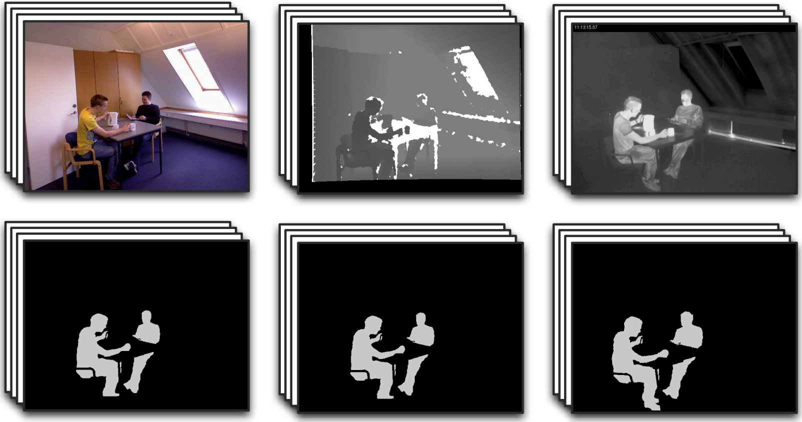
Outline

Tri-modal dataset

- Novel registered multi-modal dataset
- RGB - Depth - Thermal modalities
- 3 different scenarios
- 3 continuous image sequences
- More than 2,000 frames per sequence
- RGB - Depth pixel-level registration
- Thermal near pixel-level registration
- Manually annotated ground truth
- Registration algorithm provided

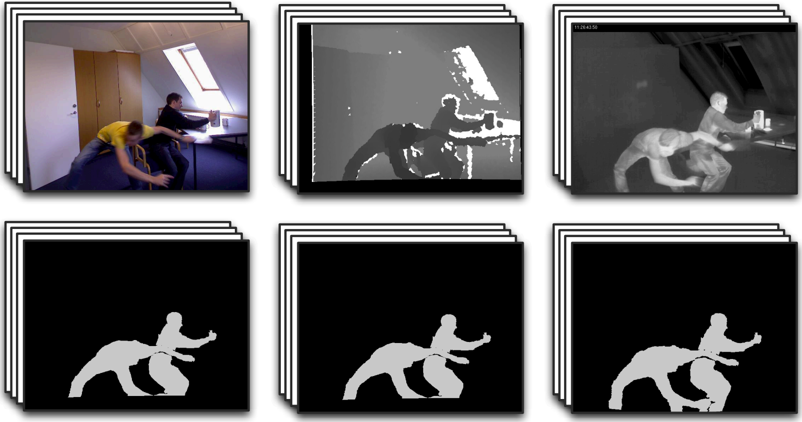
Tri-modal dataset

Scene 1



Tri-modal dataset

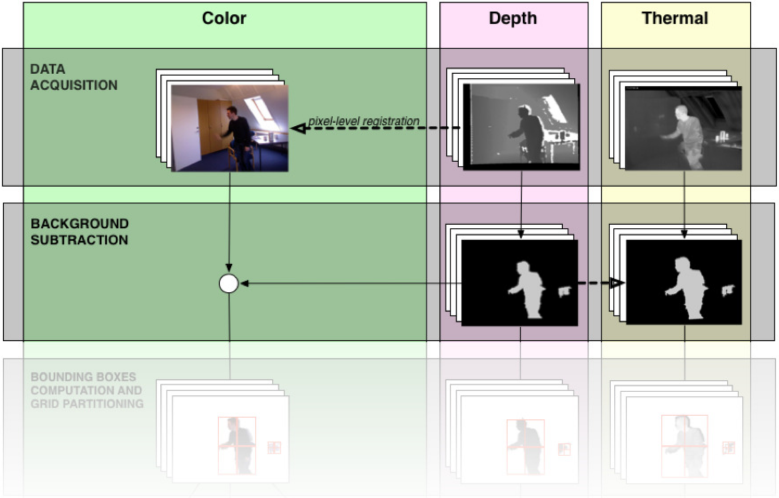
Scene 2



Tri-modal dataset
Scene 3

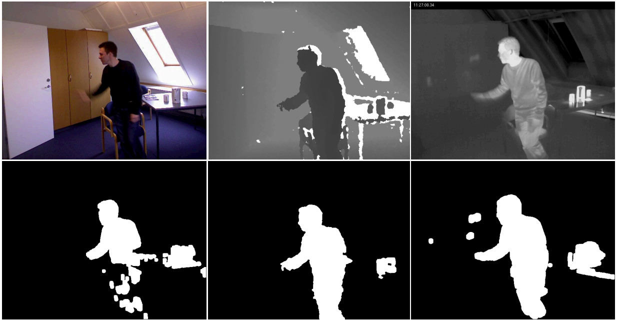
Outline

Extraction of masks



Extraction of masks
Background Subtraction

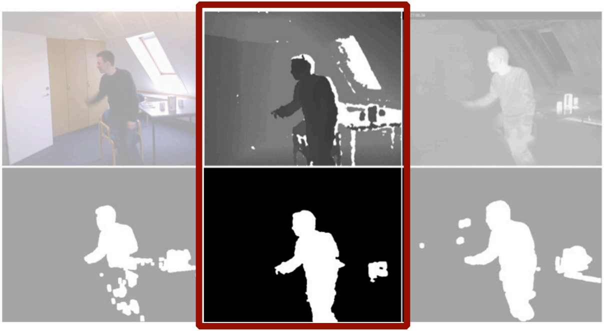
- Limit the search space.
- Non-adaptive background modeling using Mixture of Gaussians.
- Select modality:



Extraction of masks

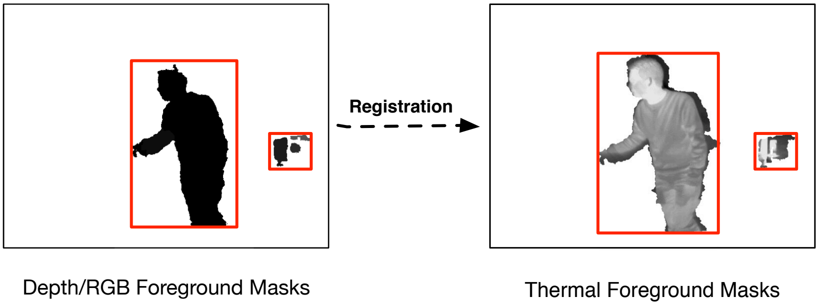
Background Subtraction

- Limit the search space.
- Non-adaptive background modeling using Mixture of Gaussians.
- Select modality: depth.

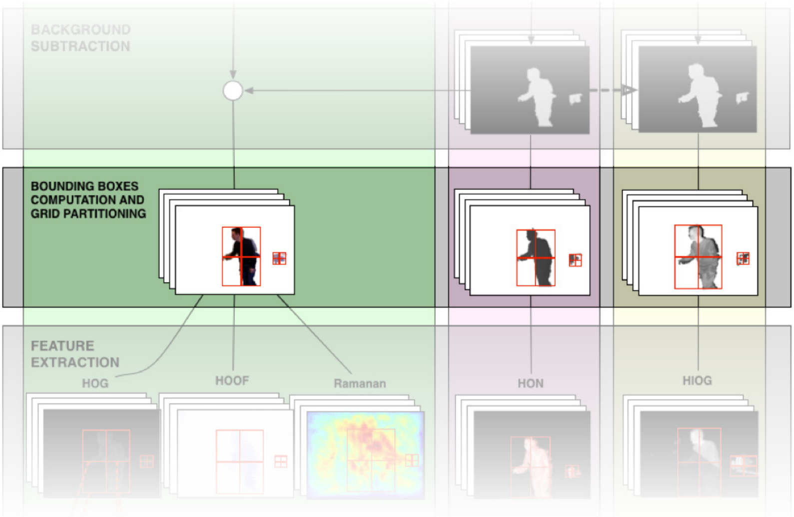


Extraction of masks

Mask registration



Extraction of regions of interest

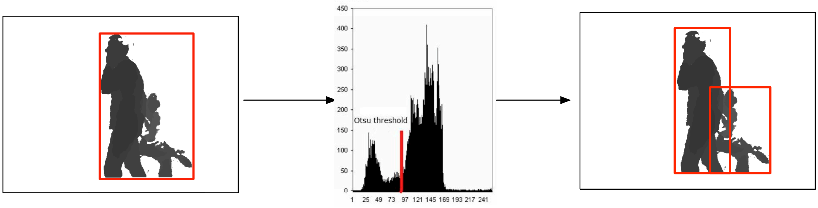


Extraction of regions of interest

Bounding box generation from regions of interest

People overlap:

- Bimodal disparity distribution.
- Otsu's threshold to split regions.



Extraction of regions of interest

Bounding box transformation and correspondence to other modalities

All modalities must have the same number of bounding boxes, corresponding to the same regions of interest.

Tasks:

1. Find correspondence between rgb/depth and thermal regions of interest.
2. Compute the corresponding bounding boxes in thermal modality generated after applying Otsu's threshold in depth modality.

Extraction of regions of interest

Bounding box transformation and correspondence to other modalities

1. Find correspondence between rgb/depth and thermal regions of interest.
    - Iterative search among depth and thermal modalities.
    - Takes into account deviation among them.
    - Best match: bounding box coordinates, amount of overlap and area similarity.
    - Correspondence function:

$$b_{iq}^{\text{thermal}} = \beta(b_{ij}^{\text{depth}}) \tag{1}$$

where b_{ij} is the j-th bounding box in frame i.

Extraction of regions of interest

Bounding box transformation and correspondence to other modalities

2 Compute the corresponding bounding boxes in thermal modality generated after applying Otsu's threshold in depth modality.

- Assuming bounding boxes of both rgb/depth and thermal modalities are proportional, find the equivalence ratio to create the split bounding boxes in thermal.
- Ratio k:

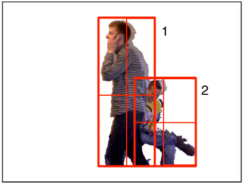
$$k_{\mathsf{h}} = \frac{h_{b_{ij}^{\mathrm{depth}}}}{h_{b_{iq}^{\mathrm{thermal}}}}, k_{\mathsf{w}} = \frac{w_{b_{ij}^{\mathrm{depth}}}}{w_{b_{iq}^{\mathrm{thermal}}}} \qquad (2)$$

where h and w are the size of a given bounding box.

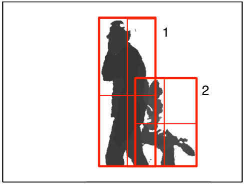
Extraction of regions of interest
Bounding box transformation and correspondence to other modalities

Result:

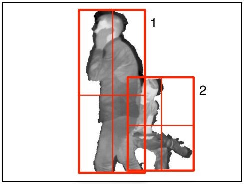
- Correspondence of regions of interest among modalities.
- Grid partitioning 2×2 cells per bounding box.



RGB



Depth



Thermal

Extraction of regions of interest

Bounding box ground truth generation

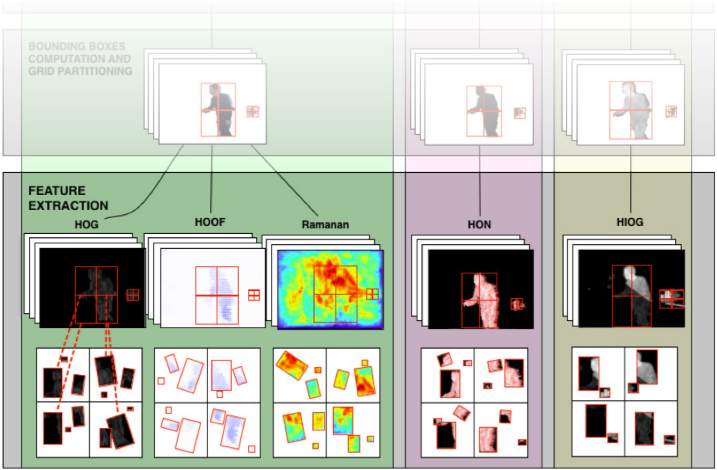
Comparing overlap between:

- Bounding boxes extracted from Ground Truth Masks
- Bounding boxes extracted from Background Subtraction Masks

Label:

$$t_r^d = \begin{cases} 0 & \text{(Object)} & \text{if overlap} \leq 0.1 \\ -1 & \text{(Unknown)} & \text{if } 0.1 < \text{overlap} < 0.6 \\ 1 & \text{(Subject)} & \text{if overlap} \geq 0.6 \end{cases}$$

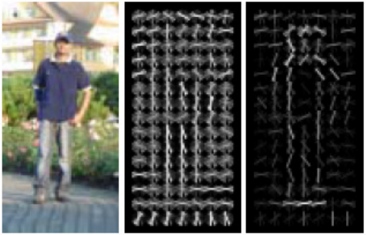
Feature extraction



Feature extraction: Color

Histogram of Oriented Gradients (HOG)

- Unsigned gradients (0 - 180 degrees).
- 9-bin histogram.
- Contribution to the histogram given by the vector magnitude.
- No block overlap applied.
- Final vector of 288 values per cell.



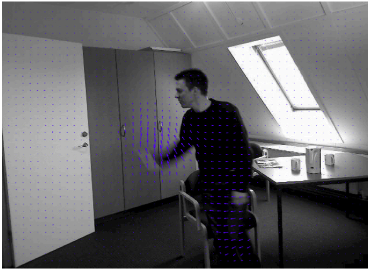
HOG

Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 886–893.

Feature extraction: Color

Histogram of Oriented Optical Flow (HOOF)

- Dense optical flow computation.
- 8-bin histogram.
- Signed gradients (0 - 360 degrees).
- Contribution to the histogram given by the vector magnitude.
- Final vector of 8 values per cell.

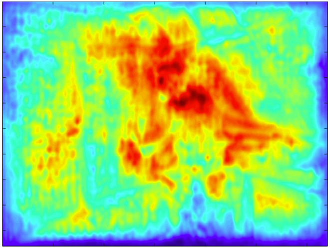


Optical flow

Feature extraction: Color

Score Maps (SM)

- Score map based on Gabor filters.
- $C = 6$ component filters per body part.
- $M = 26$ body parts.
- L scales per image.



Score maps from Ramanan et al

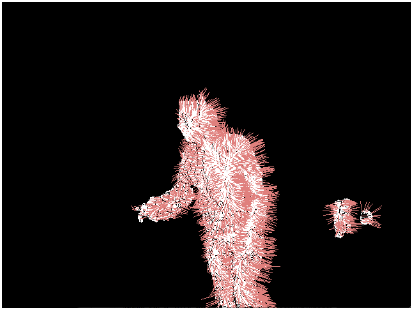
$$score(p_l) = \frac{1}{C}\frac{1}{M}\sum_{c \in C}\sum_{m \in M} score(p_l)_c^m \tag{3}$$

$$score(p) = \frac{1}{L}\sum_{l \in L} score(p_l)' \tag{4}$$

Yi Yang and Deva Ramanan. "Articulated pose estimation with flexible mixtures-of-parts". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 1385–1392.

Feature extraction: Depth
Histogram of Oriented Depth Normals (HON)

- Depth dense maps to 3D point cloud structures.
- Surface normals computations.
- Angle distribution quantized in 8-bin histogram.



Depth normals

Feature extraction: Thermal

Histogram of Thermal Intensities and Oriented Gradients (HIOG)

- Concatenation of 2 histograms:
  1. Thermal intensitities $[0, 255]$.
  2. Orientation of thermal gradients (similar to HOG).
- 8 bins per histogram.

Thermal intensities and oriented gradients

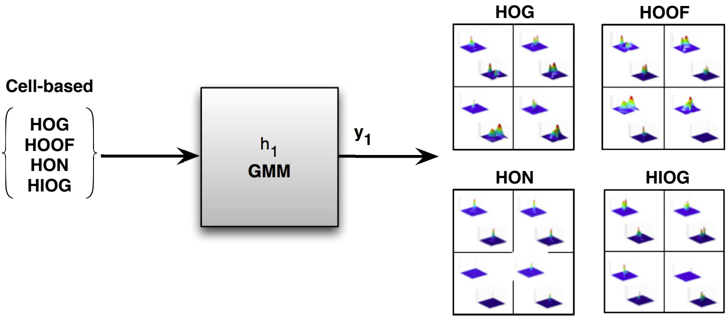
Classifiers overview

1. Statistical Learning:
    - Gaussian Mixture Models (GMM)
    - Subject and object probabilities
    - Individual prediction
2. Multi-modal fusion approaches:
    - Naive approach
    - Discriminative classifiers
    - Stacked learning fashion

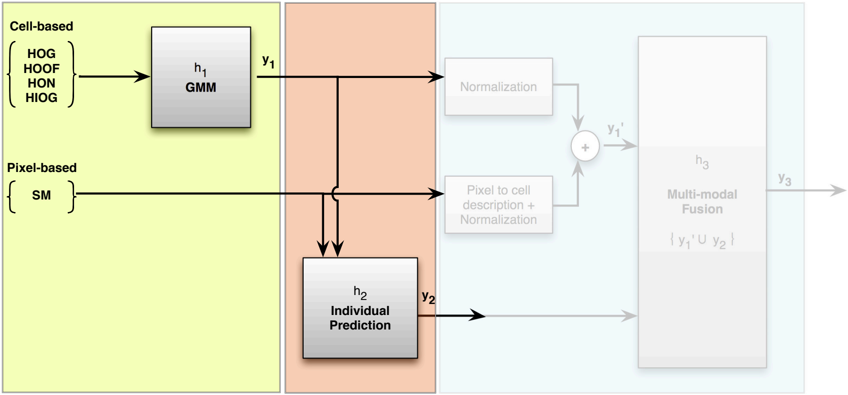
Cell classification
Gaussian Mixture Models

- Unsupervised learning method for fitting multiple Gaussians to a set of multi-dimensional data points to obtain a likelihood $\mathcal{L}$.
- Trained using Expectation Maximization algorithm.

$$\mathcal{L} = \prod_{\mathbf{x} \in \mathbf{X}}^{N} \prod_{k=1}^{K} p(\mathbf{x}|k)P(k) \qquad (5)$$



Individual Prediction



Individual Prediction

Cell-based

Predict if a region corresponds to subject or object, for each cell-based descriptor individually. Grid cell voting v:

$$v = \sum_{i,j} \mathbb{1}\{\mathcal{L}_{ij}^{d,\text{sub}} > \mathcal{L}_{ij}^{d,\text{obj}}\} \tag{6}$$

Based on a threshold v_{thr} that defines the minimum number of positive votes needed to assign the subject label to the given region:

$$v_{thr} = \frac{v_{\text{grid}} h_{\text{grid}}}{2} \tag{7}$$

Final decision $\hat{t}_r^d$:

$$\hat{t}_r^d = \mathbb{1}\left\{v > v_{thr}\right\} \bigvee \left\{\mathbb{1}\{v = v_{thr}\} \cdot \mathbb{1}\{\sum_{i,j} \left(\hat{\mathcal{L}}_{ij}^{d,\text{sub}} - \hat{\mathcal{L}}_{ij}^{d,\text{obj}}\right) > 0\}\right\} \tag{8}$$

Individual Prediction

Pixel-based

Prediction of a given region defined by:

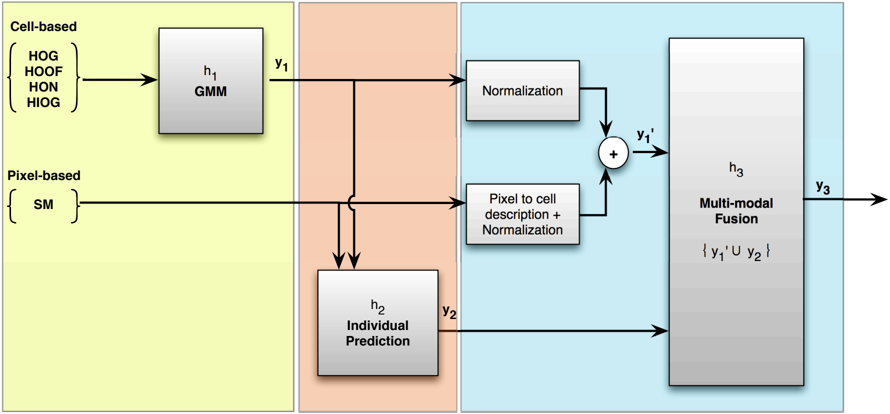
- α: minimum score of a pixel to be considered as a person.
- η: minimum percentage of pixels inside a region considered as person that are needed to label the whole region as a person.

Final decision $\hat{t}_r^d$:

$$\hat{t}_r^d = \mathbb{1}\left\{\frac{1}{N_r}\sum_{i=1}^{N_r}\mathbb{1}\{\text{score}(p_i) > \alpha\} > \eta\right\} \tag{9}$$

where N_r denotes the number of pixels of a region.

Multi-modal fusion



Multi-modal fusion

Naive approach

1. Voting among all descriptors using individual predictions $\hat{t}_r^d$.
2. If there is a strong agreement between descriptions, those descriptions that differ are not taking into account in the third step.
3. Cell level fusion:

$$\bar{\mathcal{L}}_{ij}^{d,\text{sub}} = \sum_{d \in \mathcal{D}'} \hat{\mathcal{L}}_{ij}^{d,\text{sub}}, \ \bar{\mathcal{L}}_{ij}^{d,\text{obj}} = \sum_{d \in \mathcal{D}'} \hat{\mathcal{L}}_{ij}^{d,\text{obj}} \tag{10}$$

4. Predict $\hat{t}_r$ following the same procedure as in individual prediction.

Multi-modal fusion

SVM-based approach

- Discriminative supervised binary classifier that learns a model which represents the instances as points in space, mapped in such a way that instances of different classes are separated by a hyperplane in a high dimensional space.

- Approaches:
  - Simple: $\{\hat{\mathcal{L}}_{ij}^{d,\mathrm{sub}}, \hat{\mathcal{L}}_{ij}^{d,\mathrm{obj}}\}$
  - Stacked: $\{\hat{\mathcal{L}}_{ij}^{d,\mathrm{sub}}, \hat{\mathcal{L}}_{ij}^{d,\mathrm{obj}}, \hat{t}_r^d\}$

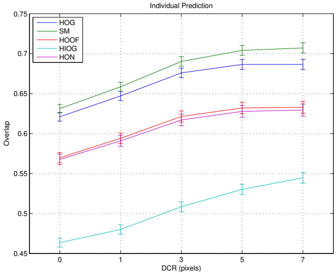
Outline

Experimental methodology and validation measures

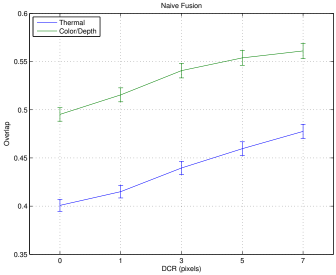
- 10-fold cross validation.
- Grid search to optimize all the parameters.
- Training without *unknown* labels.
- GMM with 3 components per Gaussian.
- SVM approaches for multi-modal fusion (simple and stacked):
    - Linear
    - RBF
- Don't care region.
- Segmentation accuracy measure: Jaccard Index

$$overlap(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{11}$$

Quantitative results

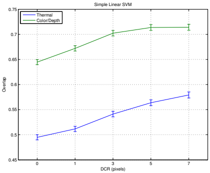


(a) Individual prediction

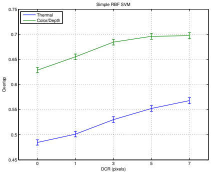


(b) Naive fusion

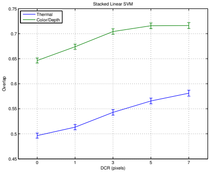
Quantitative results



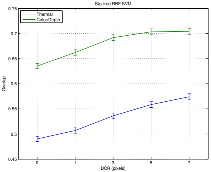
(c) Fusion using Simple linear SVM



(d) Fusion using Simple RBF SVM



(e) Fusion using Stacked Linear SVM



(f) Fusion using Stacked RBF SVM

Quantitative results

Table: Overlap results of the individual predictions for each description

DCR	HOG	SM	HOOF	HIOG	HON
0	62.10 %	63.12 %	56.97 %	46.35 %	56.76 %
1	64.71 %	65.85 %	59.41 %	47.99 %	59.09 %
3	67.59 %	69.02 %	62.13 %	50.85 %	61.70 %
5	68.65 %	70.40 %	63.20 %	53.02 %	62.77 %
7	68.65 %	70.72 %	63.28 %	54.45 %	62.94 %

Table: Overlap results of fusion using Stacked Linear SVM for each modality

DCR	Thermal	Color/Depth
0	49.64 %	64.65 %
1	51.33 %	67.39 %
3	54.29 %	70.43 %
5	56.56 %	71.58 %
7	58.11 %	71.63 %

Qualitative results

Comparison between masks generated after background subtraction and masks generated using Stacked Linear SVM.

Outline

Conclusions

- A solution for human body segmentation in multi-modal data has been proposed.
- A novel tri-modal dataset has been presented, containing RGB - Depth - Thermal modalities.
- Results show variable performance for the different modalities when segmenting people in multi-modal data, and improved segmentation accuracy of the multi-modal GMM-SVM stacked learning method.

Future work

- Impact factor journal in progress in collaboration with Aalborg University and HuPBA group.
- Silhouette masks refinement using Grabcuts.
- If pixel-level registration available among all modalities:
  - Combination of different modalities in background subtraction.
  - Pixel-level feature extraction.
  - Pixel-level description.
- Extensive validation in real surveillance scenarios as a first real case study, including gesture recognition methodologies (planning just started with Aalborg University).

Thank you.