SEMANTIC FACE SEGMENTATION FROM VIDEO STREAMS IN THE WILD

Student: Deividas Skiparis
Supervisors: Pascal Landry (Imersivo) and Sergio Escalera (UB)

imersivo

UNIVERSITAT ROVIRA i VIRGILI

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

UNIVERSITAT DE BARCELONA

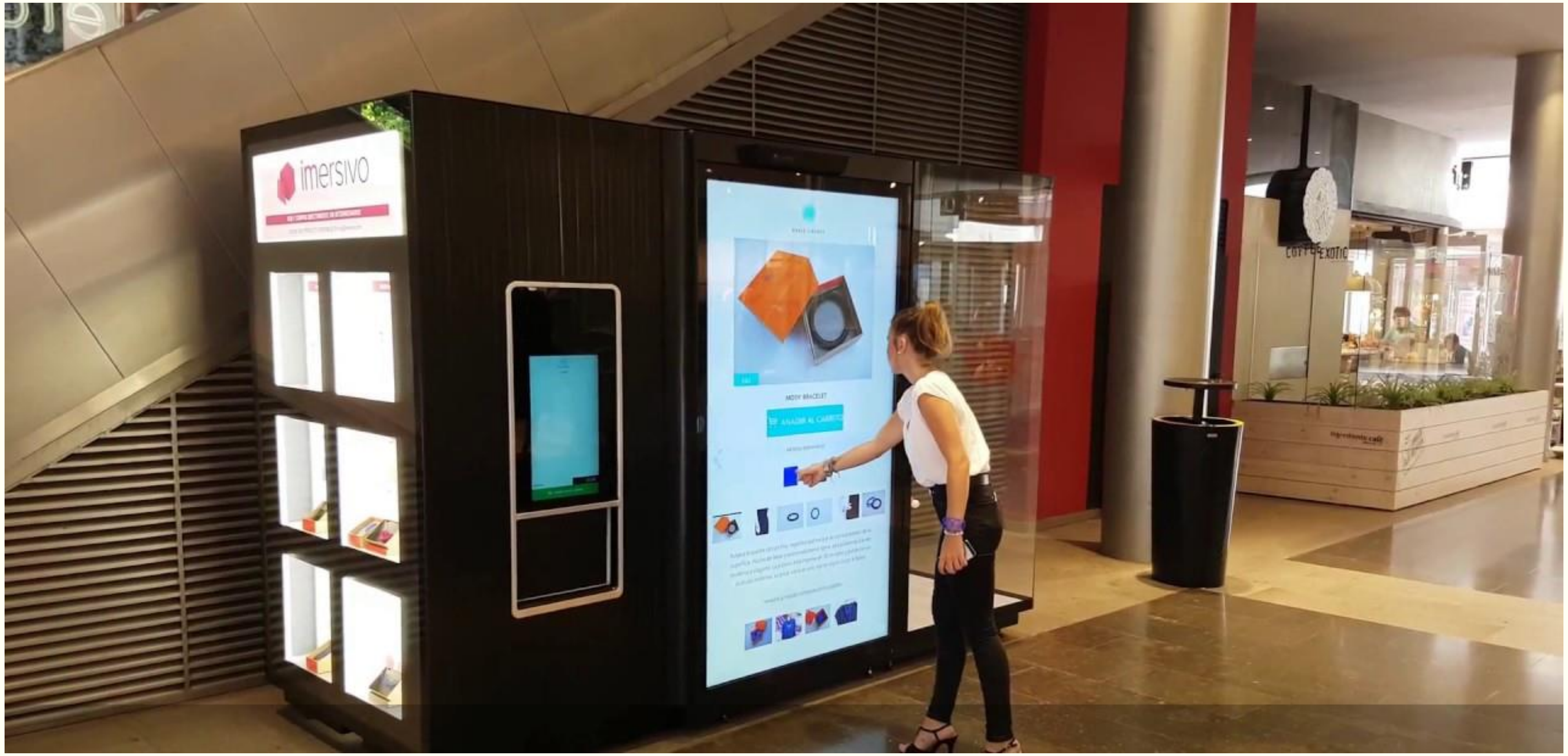
In Collaboration With Imersivo SL

imersivo

Problem

- Fashion Expert Functionality
- Needs to know Skin, Hair and Eye colors
- No available off-the-shelf libraries for semantic face segmentation
- Requirements:
  - 1fps
  - Reasonable accuracy
  - Respects current hardware

Application



- Possible applications:
    Recommendation engines, fashion consulting, audience inspection, etc.

Proposed solution

- Based on multi-stage graphical model approach
  - 1st stage - Bi-partite graph-cut for BGD removal
  - 2nd stage – Bi-partite graph-cut for segmentation
- Temporal feature for error mitigation

Related Work

- No comparable work exists
- Collection of already existing methods
- Necessary components
  - Face detection
  - Landmark fitting
  - Segmentation
  - Video segmentation

Related Work – Face Detection

- Vital for segmentation to localize the ROI
- Viola&Jones – integral image and Haar-like features [1]

Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features". In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Vol. 1. IEEE. 2001, pp. I–I

Related Work – Face Detection

- Vital for segmentation to localize the ROI
- Viola&Jones – integral image and Haar-like features
- Deep face detectors

Henry A Rowley, Shumeet Baluja, and Takeo Kanade. "Neural network-based face detection". In: *IEEE Transactions on pattern analysis and machine intelligence* 20.1 (1998), pp. 23–38

Christophe Garcia and Manolis Delakis. "A neural architecture for fast and robust face detection". In: *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. Vol. 2. IEEE. 2002, pp. 44–47

Related Work – Face Detection

- Vital for segmentation to localize the ROI
- Viola&Jones – integral image and Haar-like features
- Deep face detectors
- Deformable Part Models

Pedro F Felzenszwalb et al. "Object detection with discriminatively trained part-based models". In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2010), pp. 1627–1645

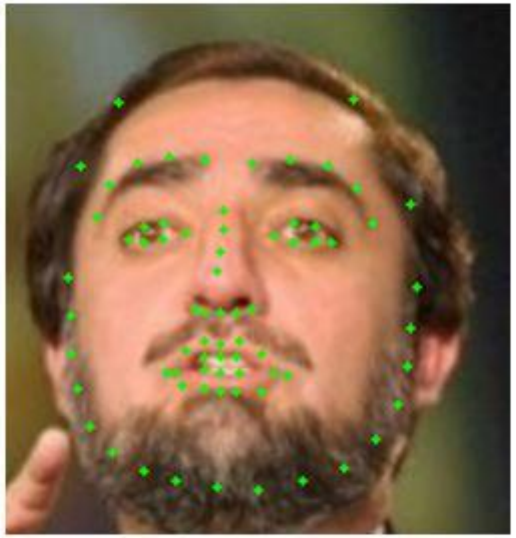
Related Work – Face Detection

- Vital for segmentation to localize the ROI
- Viola&Jones – integral image and Haar-like features
- Deep face detectors
- Deformable Part Models
- 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.

Related Work – Landmark Fitting

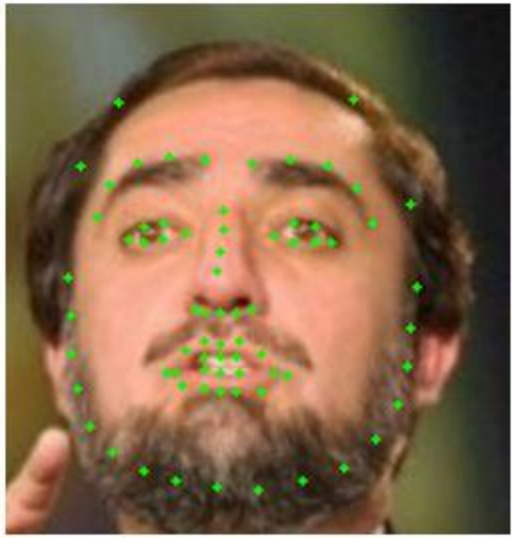
- Supervised gradient descent



Xuehan Xiong and Fernando De la Torre. "Supervised descent method and its applications to face alignment". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 532–539

Related Work – Landmark Fitting

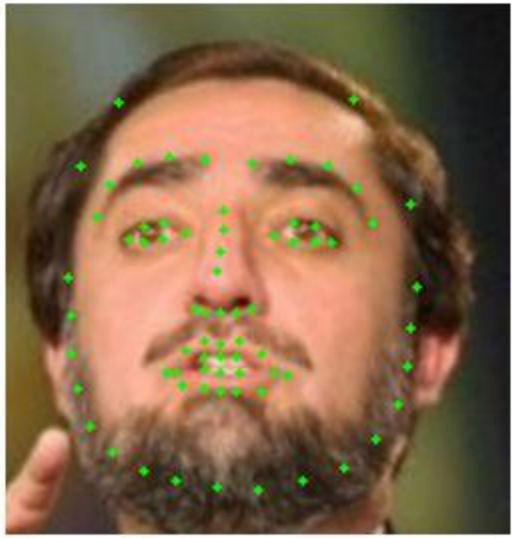
- Supervised gradient descent
- Ensemble of regression trees



Vahid Kazemi and Josephine Sullivan. "One millisecond face alignment with
an ensemble of regression trees". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1867–1874

Related Work – Landmark Fitting

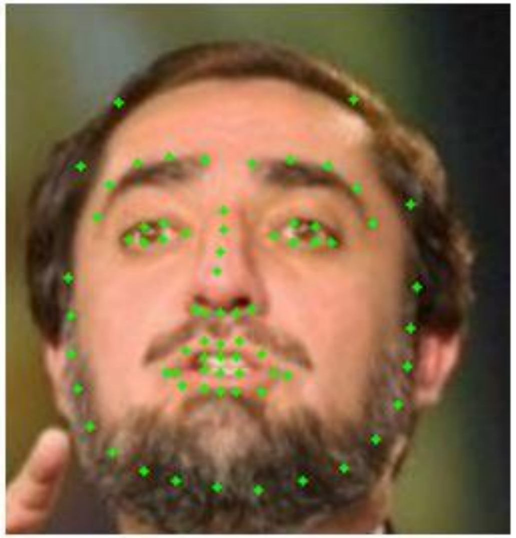
- Supervised gradient descent
- Ensemble of regression trees
- CNNs - broad range of poses and virtually invariant to occlusions.



Zhanpeng Zhang et al. “Facial landmark detection by deep multi-task learning”. In: *European Conference on Computer Vision*. Springer. 2014, pp. 94–108

Amin Jourabloo and Xiaoming Liu. “Large-pose face alignment via CNNbased dense 3D model fitting”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4188–4196

Related Work – Landmark Fitting



- Supervised gradient descent
- Ensemble of regression trees
- CNNs - broad range of poses and virtually invariant to occlusions
- Other methods considered: Active Appearance models [1] multidimensional morphable models [2] and template tracking [3]

Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. "Active appearance models". In: *European conference on computer vision*. Springer. 1998, pp. 484–498

Michael J Jones and Tomaso Poggio. "Multidimensional morphable models". In: *Computer Vision, 1998. Sixth International Conference on*. IEEE. 1998, 683–688

Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. "Robust and efficient parametric face alignment". In: *Computer Vision (ICCV), 2011 IEEE International Conference On*. IEEE. 2011, pp. 1847–1854.

Related Work - Segmentation

- Most recent studies use ANNs
  - CRF with Adaboosted unary classifier and epitome priors

Jonathan Warrell and Simon JD Prince. "Labelfaces: Parsing facial features by multiclass labeling with an epitome prior". In: *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE. 2009, pp. 2481–2484

Related Work - Segmentation

- Most recent studies use ANNs
  - CRF with Adaboosted unary classifier and epitome priors
  - CRF with Restricted Boltzmann Machine prior

Andrew Kae et al. "Augmenting CRFs with Boltzmann machine shape priors for image labeling". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 2019–2026

Related Work - Segmentation

- Most recent studies use ANNs
  - CRF with Adaboosted unary classifier and epitome priors
  - CRF with Restricted Boltzmann Machine prior
  - DPM for hierarchical face parsing + new deep learning strategy for segmentation

Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Hierarchical face parsing via deep learning". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 2480–2487

Related Work - Segmentation

- Most recent studies use ANNs
  - CRF with Adaboosted unary classifier and epitome priors
  - CRF with Restricted Boltzmann Machine prior
  - DPM for hierarchical face parsing + new deep learning strategy for segmentation
- Depth sensors

Chenxi Zhang, Liang Wang, and Ruigang Yang. "Semantic segmentation of urban scenes using dense depth maps". In: *European Conference on Computer Vision*. Springer. 2010, pp. 708–721

Related Work - Segmentation

- Most recent studies use ANNs
  - CRF with Adaboosted unary classifier and epitome priors
  - CRF with Restricted Boltzmann Machine prior
  - DPM for hierarchical face parsing + new deep learning strategy for segmentation
- Depth sensors
- Graphical models

Akira Suga et al. "Object recognition and segmentation using SIFT and Graph Cuts". In: *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE. 2008, pp. 1–4

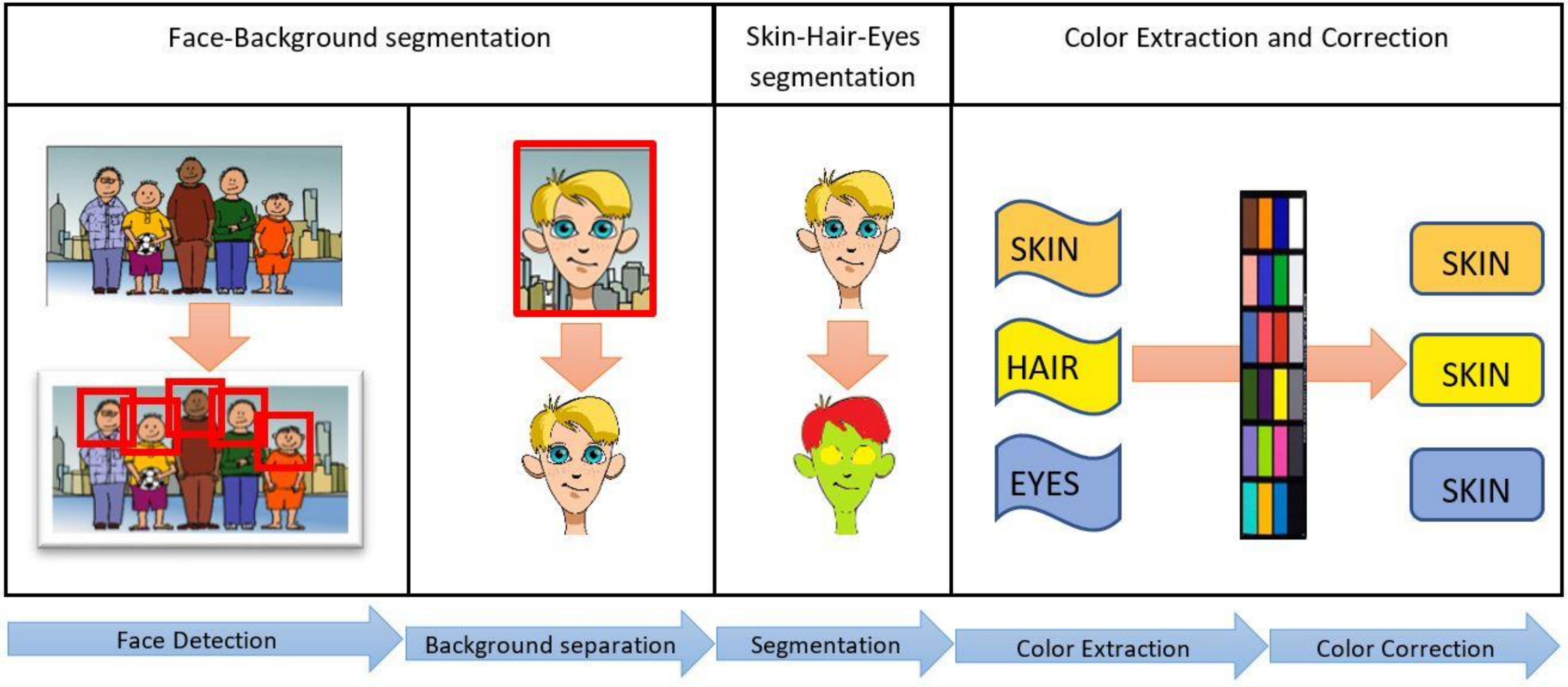
Related Work - Temporal Segmentation

- Generally accepted definition – spatio-temporal label propagation in a video (object tracking, gesture recognition)

Anestis Papazoglou and Vittorio Ferrari. "Fast object segmentation in unconstrained video". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 1777–1784

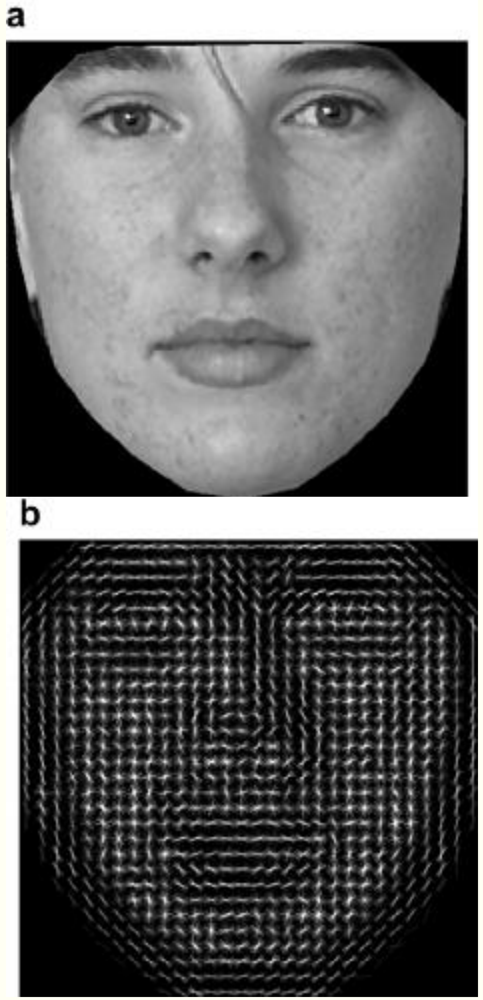
- This paper simplifies the definition of video segmentation
- Temporal segmentation as a self-validation mechanism

Proposed System Overview



Face detection and Landmark Fitting

- Face detector: HOG
  - The basic idea - local appearance and shape can be characterized by the distribution of local intensity gradients (edge directions)
  - Divide image window into small spatial regions ("cells"),
  - For each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the contrast-normalize (invariance to illumination) pixels of the cell.
  - Tile the detection window with a dense (in fact, overlapping) grid of HOG descriptors and use the combined feature vector for classification



Face detection and Landmark Fitting

- Face detector: HOG
- Facial Landmarks: Ensemble of regression trees
  - Predictive model
  - Composed of a weighted combination of multiple regression trees
  - Each regressor is learned using gradient boosting tree algorithm and square error loss
  - Combining multiple regression trees increases predictive performance
- Facial landmarks serve multiple purposes:
  - Help to understand the face pose – reject problematic poses
  - Perform eye-area segmentation
  - Precise location of facial parts – used for color modeling in later stages.

Algorithm 1 Learning r_t in the cascade

Have training data $\{(I_{\pi_i}, \hat{\mathbf{S}}_i^{(t)}, \Delta\mathbf{S}_i^{(t)})\}_{i=1}^N$ and the learning rate (shrinkage factor) $0 < \nu < 1$

1. Initialise

$$f_0(I, \hat{\mathbf{S}}^{(t)}) = \arg\min_{\gamma \in \mathbb{R}^{2p}} \sum_{i=1}^{N} \|\Delta\mathbf{S}_i^{(t)} - \gamma\|^2$$

2. for $k = 1, \ldots, K$:

   (a) Set for $i = 1, \ldots, N$

$$\mathbf{r}_{ik} = \Delta\mathbf{S}_i^{(t)} - f_{k-1}(I_{\pi_i}, \hat{\mathbf{S}}_i^{(t)})$$

   (b) Fit a regression tree to the targets $\mathbf{r}_{ik}$ giving a weak regression function $g_k(I, \hat{\mathbf{S}}^{(t)})$.

   (c) Update

$$f_k(I, \hat{\mathbf{S}}^{(t)}) = f_{k-1}(I, \hat{\mathbf{S}}^{(t)}) + \nu\, g_k(I, \hat{\mathbf{S}}^{(t)})$$

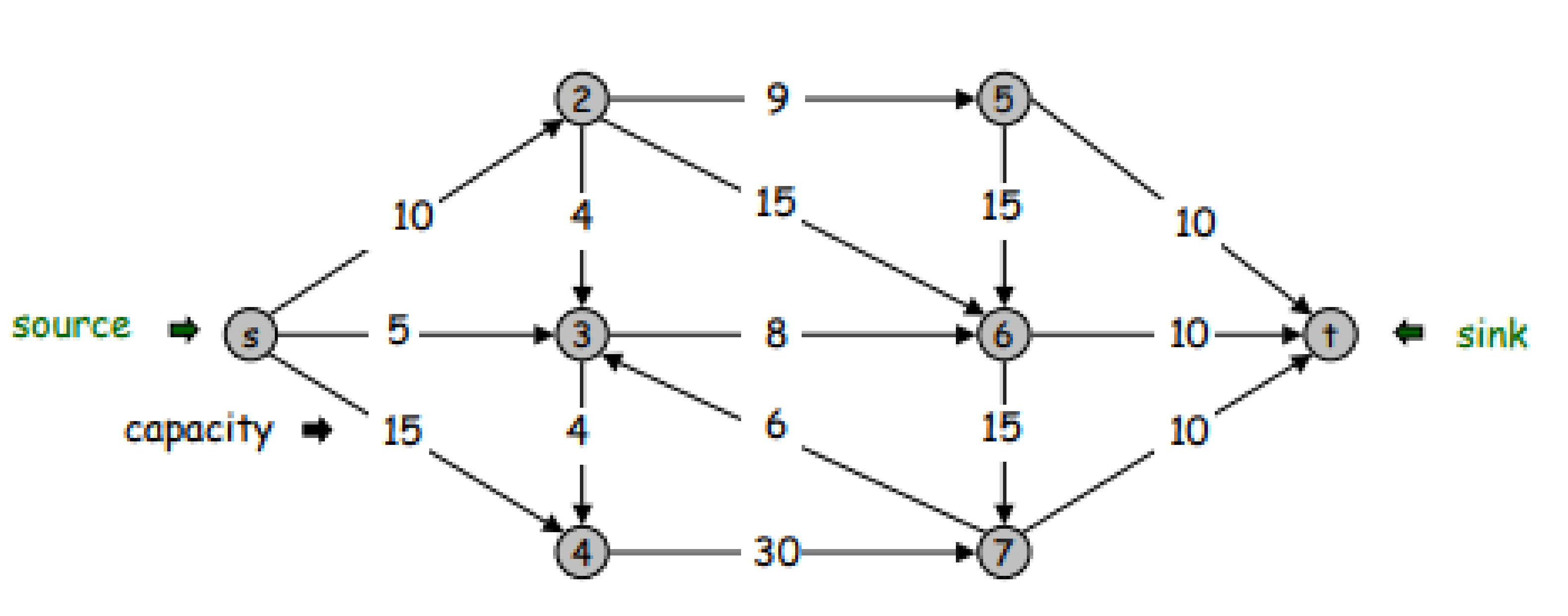
3. Output $r_t(I, \hat{\mathbf{S}}^{(t)}) = f_K(I, \hat{\mathbf{S}}^{(t)})$

Background removal

- Uses bounding rectangles derived from Facial Landmarks
- GMM for each ROI
- Min-Cut Max-Flow graph cut

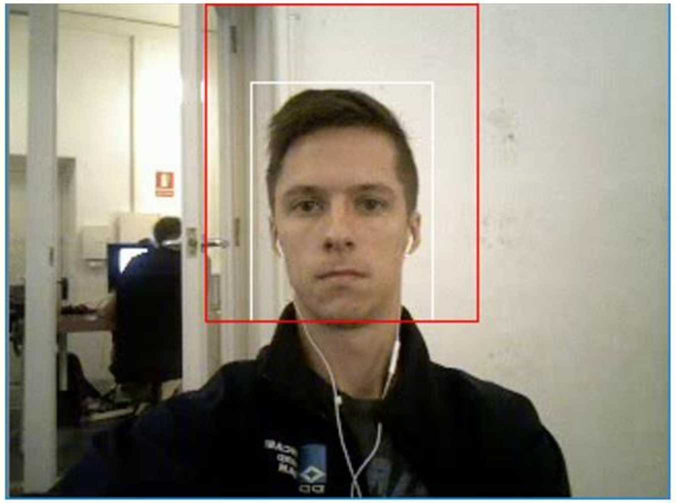
$$\text{maximize } |f| = \nabla_s$$

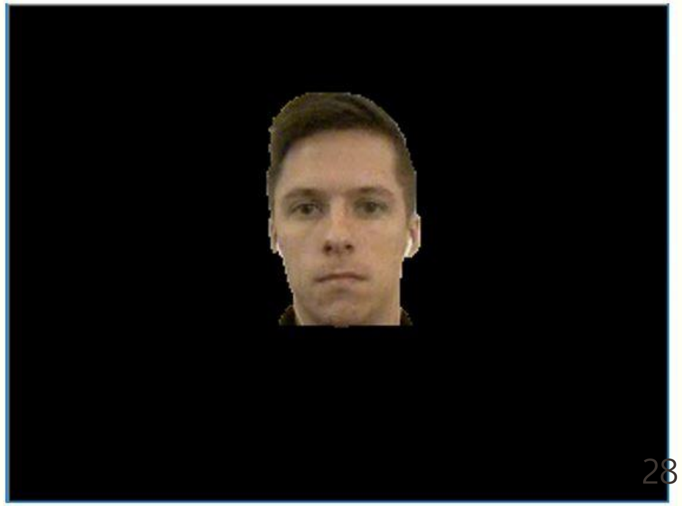
$$\text{minimize } \sum_{(i,j)\in E} c_{ij} d_{ij}$$



Background removal

- Uses bounding rectangles derived from Facial Landmarks
- GMM for each ROI
- Min-Cut Max-Flow graph cut
- OPENCV implementation – GrabCut





Skin-Hair-Eyes Segmentation

- Skin-Hair – Graph Cut with α-expansion minimization algorithm

- 1. Start with any labeling
- 2. run through all labels and for each label a
  - 2a. compute optimal a-expansion move
  - 2b. if better energy found, accept the move
- 3. Stop if no label change, otherwise go to 2

$$E(f) = \sum_{p \in P} D_p(i_p, f_p) + \sum_{p,q \in N} V_{p,q}(f_p, f_q)$$

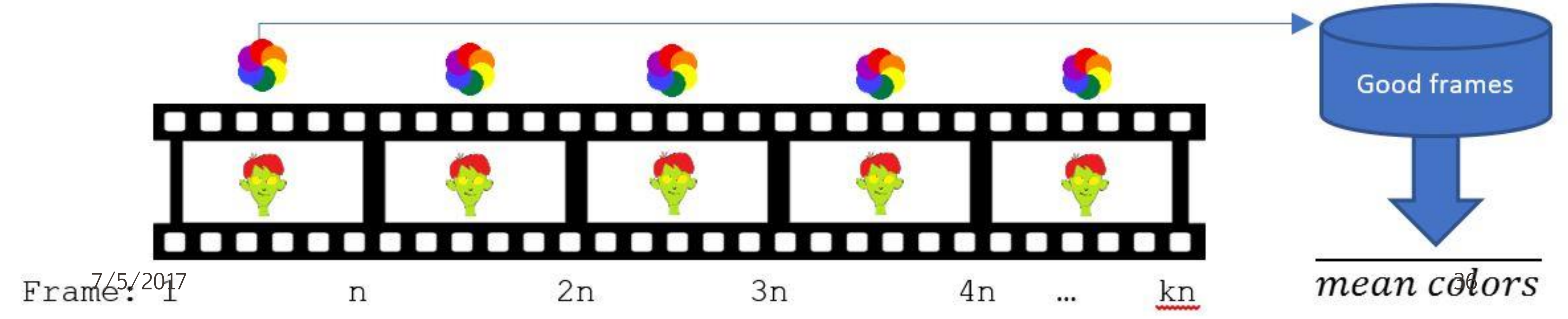
- Multi-class graph cut algorithm was chosen in order to extend applicability

Color Extraction

- Build normalized color histograms of N bins per channel for each semantic area
- Weighted average of X largest bins was calculated to get a final color for a semantic category
- Optimal X is later found experimentally

Temporal Segmentation

- Faces in the wild are often noisy
- Semantic areas' colors are propagated through frames
- Procedure:
  - Frame every 2sec is segmented
  - Color is obtained for each semantic category
  - The extracted colors are saved into a time series vector.
  - The extracted colors are compared to colors in two previous time steps (current-1 and current-2).
  - If the results are within a pre-determined threshold, the values are saved as



Experiments - Data

- YouTube personality dataset
- ChaLearn Looking at People Workshop on Automatic Personality Analysis and First Impressions Challenge @ ECCV2016
- 100 Random videos (0.9/0.1 training/validation)

Gender	Age	Skin	Accessories	Hair Amount	Hair	Facial Hair
Female (59)	20s (44)	Dark (15)	Glasses (7)	Little (10)	Dark (74)	Heavy (4)
Male (41)	30s (37)	Light (85)	Hat (10)	None (6)	Light (14)	Little (14)
	40s (15)		Head-band (1)	Normal (84)	Mixed (2)	None (77)
	50s (4)		None (82)		None (6)	Normal (5)
					Red (4)	

Optimal Parameter Search

- Modular pipeline → Many tunable parameters (21)
- 9 selected as having the most importance

Group1	Group2	Group3,4,5
Unary	hair_sample_size	top_N_colors_C
Pair-wise	hair_TH	TH_C
SP_amount	hist_bin	
	hist_bins_2	

- Number of iterations 432×10^6 to a much more manageable 640

Optimal Parameter Search

No	System Part	Group	Parameter	Best Values
1	Face Segmentation	Group1	Unary weight	9
2			Pair-wise weight	31
3			SP_amount	350
4		Group2	hair_sample_size	0.15
5			hair_TH	0.35
6	Color Extraction		hist_bin	16
7			hist_bins_2	128
8		Group3	top_N_colors_{C}	4
9	Temporal Analysis		TH_{C}	50

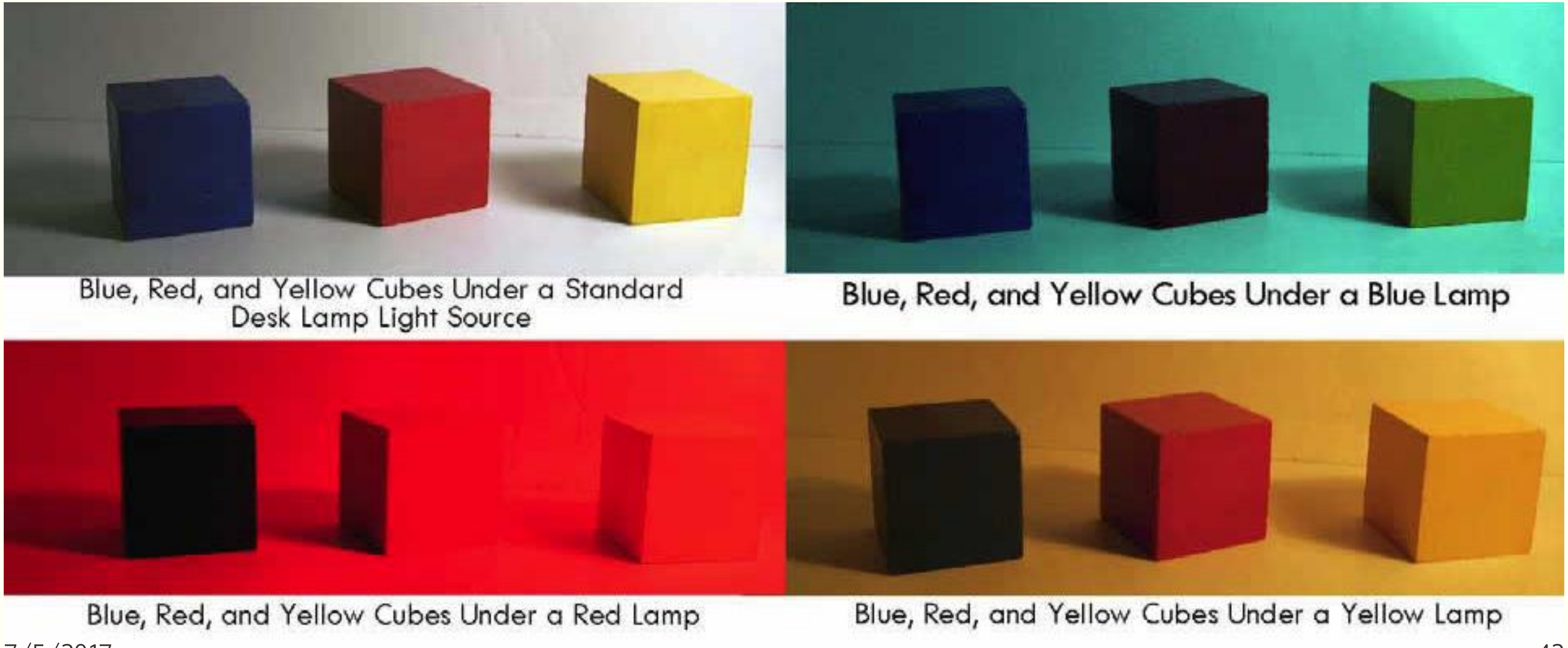
Results – Color Extraction

- All measures are calculated over the validation set
- Ground truth values have been marked manually on all videos

	Overall
mErr	31.92
Stddev	26.58
% error	11.03

Color Correction

- Light has high influence on perceived colors



Blue, Red, and Yellow Cubes Under a Standard Desk Lamp Light Source

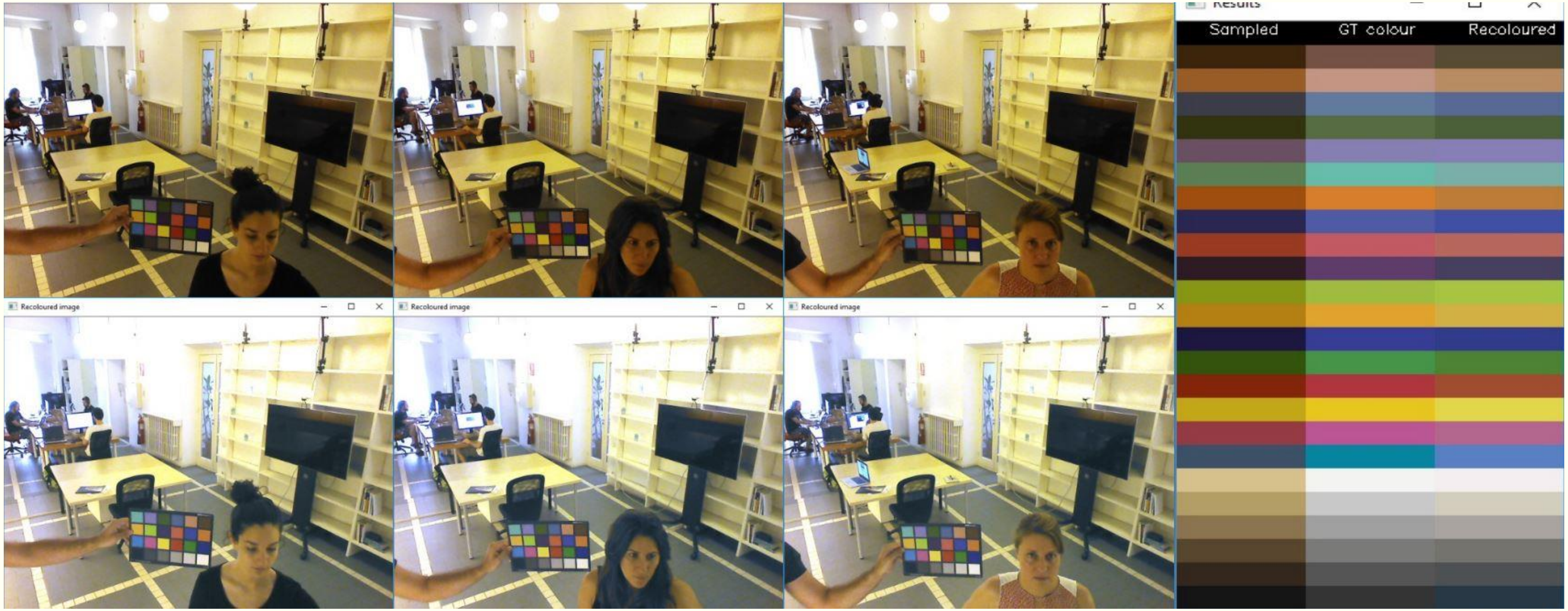
Blue, Red, and Yellow Cubes Under a Blue Lamp

Blue, Red, and Yellow Cubes Under a Red Lamp

Blue, Red, and Yellow Cubes Under a Yellow Lamp

Results – Color Correction

- Comprehensive testing is out of scope



Discussion - Problems

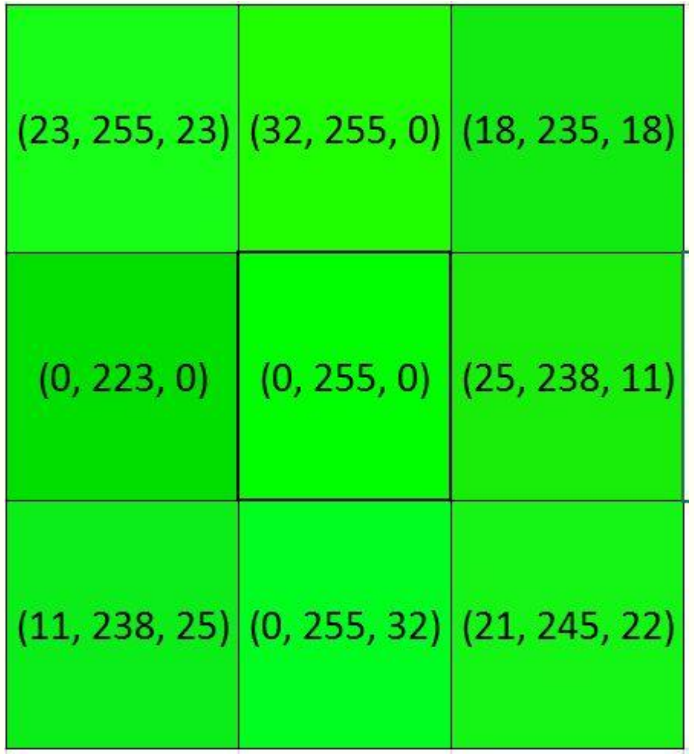
- 'Just noticeable' difference for LAB - 3.2. Mean error of the system – 32.
- StdDev - 26.6. System exhibits a lot of variance, thus the error can fluctuate significantly.
- Possible sources of error:
  - Inability to filter out face occlusions – large accessories affect final color
  - Busy and dynamic background could result in poor background - foreground segmentation
  - Bald spots
  - Lighting conditions. Strong light sources can generate shinny spots

Fail cases

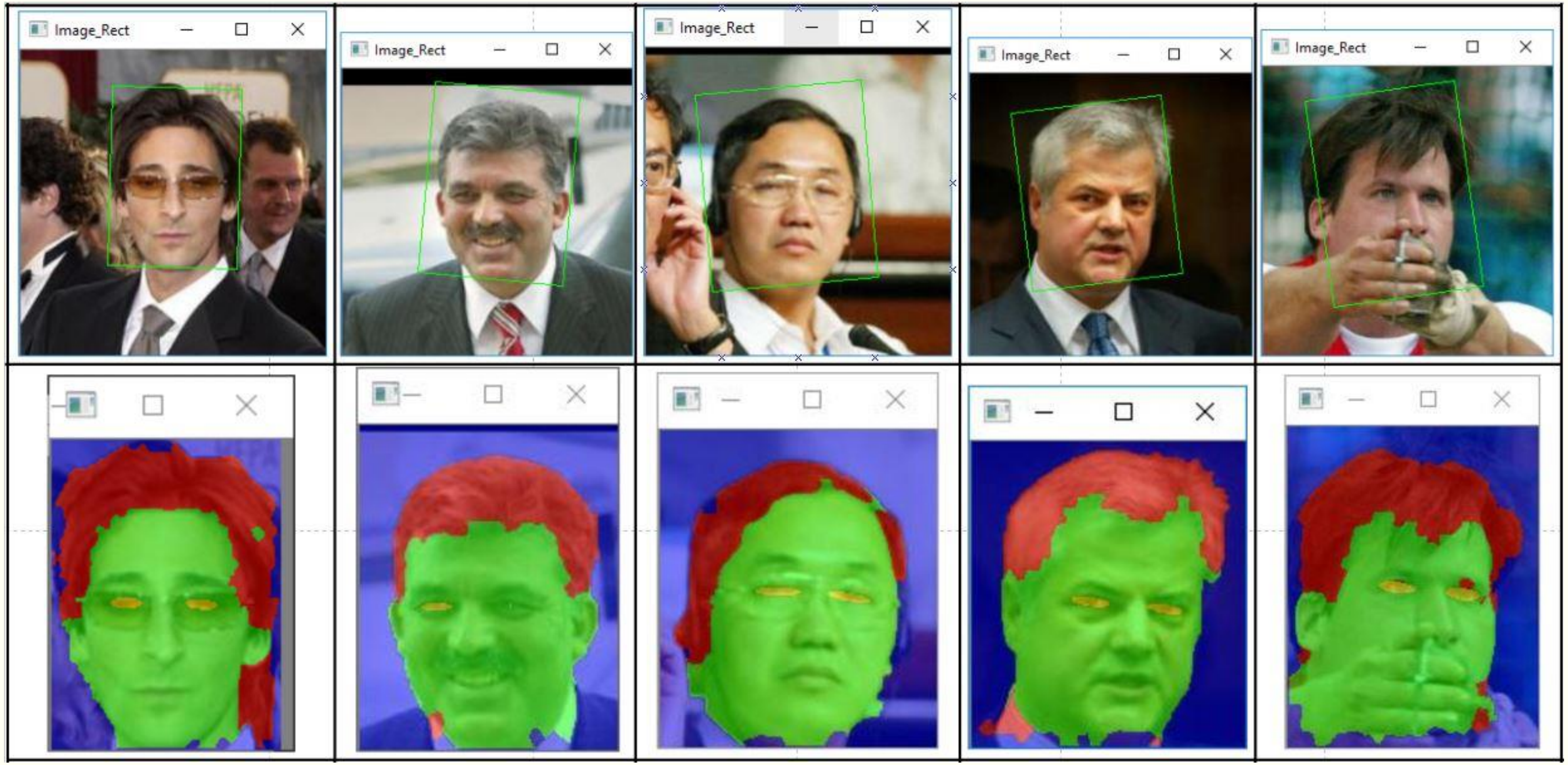
Poor color extraction due to face oclussions	Poor color extraction due to hair oclussions	Sub-standard foreground-background segmentation	Shinny spots considered as hair patches	Poor segmentation due to face shadows

Discussion – Strong points

- 5fps on a standard CPU – (1fps requirement)
- Extracted colors will be used for fashion recommendations → error = 32 (11%) is still sufficiently low.
- Perceptually negligible →



Good cases - segmentation



Adaptability

- Designed to extract colors of semantic face components, but easily adaptable to other applications
- Only need to change the object detector
- Second stage of graph cut uses multi-class alpha-expansion
- Other possible application: color extraction of currently worn clothes through segmentation of human body

Future work

- Ability to recognize occlusions
- Correctly segment bald-spots and bald people
- Accuracy of the proposed system relies largely on the quality of segmentation. But all current SOTA methods for segmentation are based on CNNs
- Introduction of CNN would significantly improve the segmentation quality, thus potentially making this color extraction system a SOTA

Questions