Motion and Activity Analysis with Spatiotemporal Local Binary Patterns

Matti Pietikäinen and Guoying Zhao
{mkp,gyzhao}@ee.oulu.fi

Machine Vision Group
University of Oulu, Finland
http://www.ee.oulu.fi/mvg/

Contents

1. Introduction to LBP operators in spatial domain
2. Motion analysis with spatiotemporal LBPs
3. Summary

Dynamic textures (R Nelson & R Polana: IUW, 1992; M Szummer & R Picard: ICIP, 1995; G Doretto et al., IJCV, 2003)

Local Binary Pattern and Contrast operators

Ojala T, Pietikäinen M & Harwood D (1996) A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29:51-59.

An example of computing LBP and C in a 3x3 neighborhood:

example

6	5	2
7	6	1
9	8	7

thresholded

1	0	0
1	▨	0
1	1	1

weights

1	2	4
128	▨	8
64	32	16

Pattern = **11110001**

LBP = 1 + 16 +32 + 64 + 128 = **241**

C = (6+7+8+9+7)/5 - (5+2+1)/3 = **4.7**

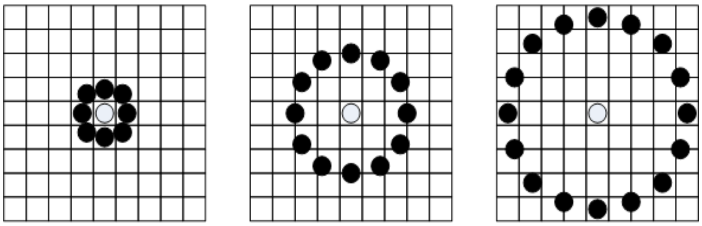
Important properties:

- LBP is invariant to any monotonic gray level change
- computational simplicity

Multiscale LBP

Ojala T, Pietikäinen M & Mäenpää T (2002) Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7):971-987.

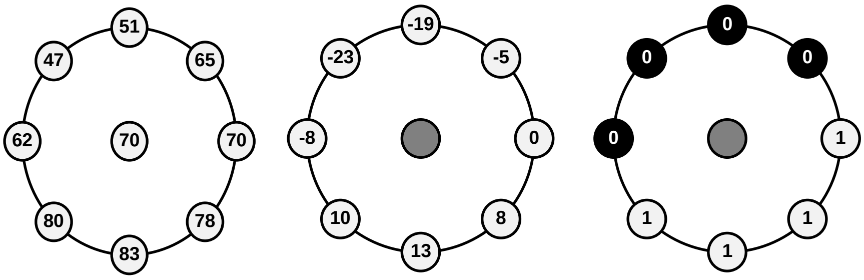
- arbitrary circular neighborhoods
- uniform patterns
- multiple scales
- rotation invariance
- gray scale variance as contrast measure



P = 8, R = 1.0 P = 12, R = 2.5 P = 16, R = 4.0

The value of the LBP code of a pixel (x_c, y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \qquad s(x) = \begin{cases} 1, if\ x \geq 0; \\ 0, otherwise. \end{cases}$$

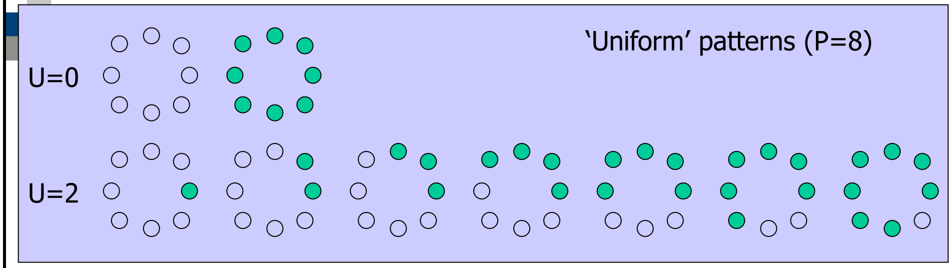


1. Sample 2. Difference 3. Threshold

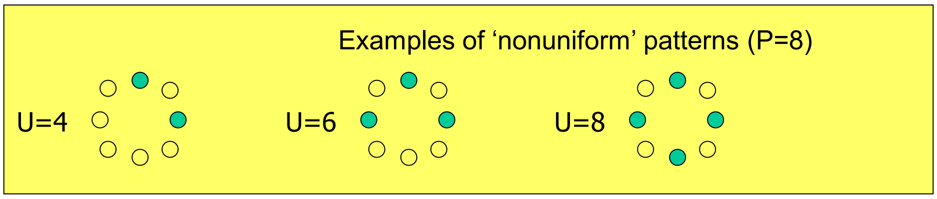
1*1 + 1*2 + 1*4 + 1*8 + 0*16 + 0*32 + 0*64 + 0*128 = 15

4. Multiply by powers of two and sum

‘Uniform’ patterns



‘Uniform’ patterns (P=8)

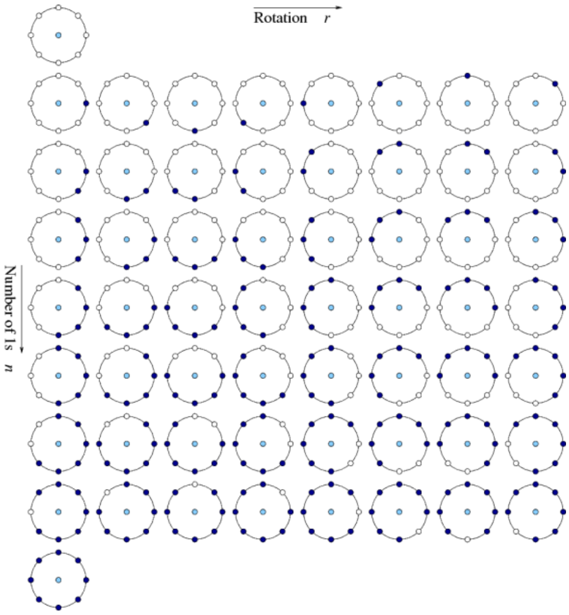


Examples of ‘nonuniform’ patterns (P=8)

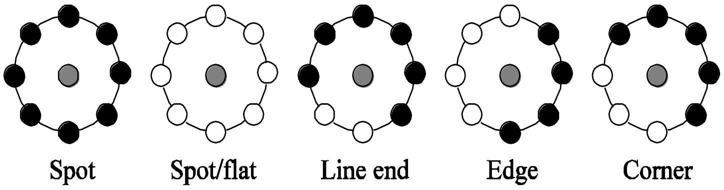
- Bit patterns with 0 or 2 transitions *0→1* or *1→0* when the pattern is considered circular

- All non-uniform patterns assigned to a single bin

- 58 uniform patterns in case of 8 sampling points

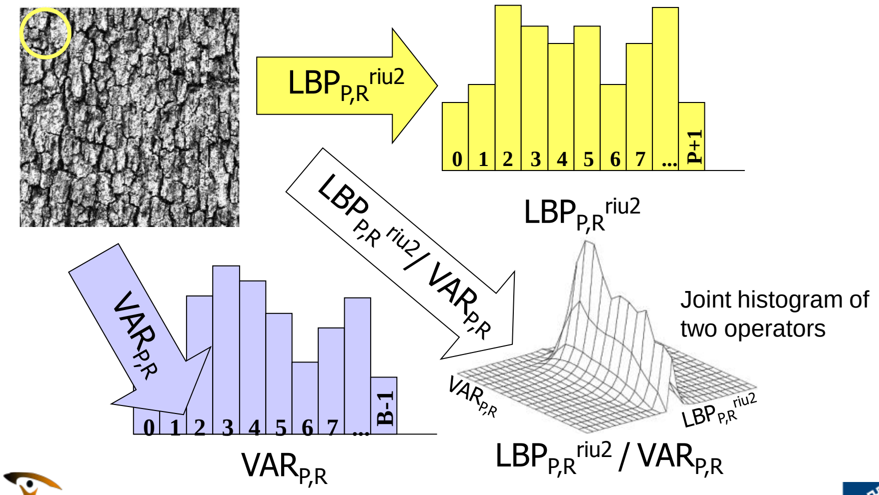


Texture primitives (“micro-textons”) detected by the uniform patterns of LBP



Estimation of empirical feature distributions

Input image (region) is scanned with the chosen operator(s), pixel by pixel, and operator outputs are accumulated into a discrete histogram



Joint histogram of two operators

Multiscale analysis

Information provided by N operators can be combined simply by summing up operatorwise similarity scores into an aggregate similarity score:

$$L_N = \sum_{n=1}^{N} L_n \quad \text{e.g.} \quad LBP_{8,1}^{riu2} + LBP_{8,3}^{riu2} + LBP_{8,5}^{riu2}$$

Effectively, the above assumes that distributions of individual operators are independent

Nonparametric classification principle

Sample S is assigned to the class of model M that maximizes

$$L(S,M) = \sum_{b=0}^{B-1} S_b \ln M_b$$

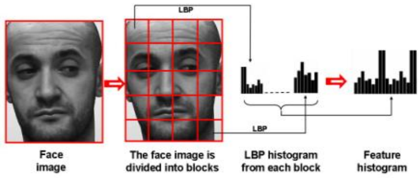
Many other dissimilarity measures can be used (chi square, histogram intersection, Kullback-Leibler divergence, Jeffrey′s divergence, etc.)

Nonparametric: no assumptions about underlying feature distributions are made!!

Face analysis using local binary patterns

- Face recognition is one of the major challenges in computer vision

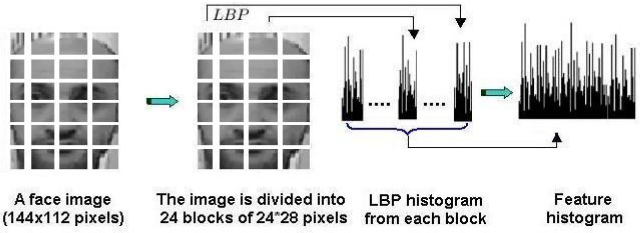
- We proposed (ECCV 2004, PAMI 2006) a face descriptor based on LBP’s
- Our method has already been adopted by many leading scientists and groups
- Computationally very simple, excellent results in face recognition and authentication, face detection, facial expression recognition, gender classification


LBP
LBP
Face image
The face image is divided into blocks
LBP histogram from each block
Feature histogram

Face description with LBP

Ahonen T, Hadid A & Pietikäinen M (2006) Face description with local binary patterns: application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12):2037-2041. (an early version published at ECCV 2004)

A facial description for face recognition:


LBP
....
....
A face image (144x112 pixels)
The image is divided into 24 blocks of 24*28 pixels
LBP histogram from each block
Feature histogram

Dynamic texture recognition

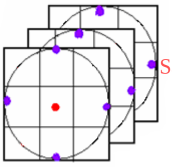
Zhao G & Pietikäinen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6):915-928. (parts of this were earlier presented at ECCV 2006 Workshop on Dynamical Vision and ICPR 2006)

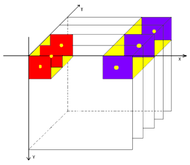
Dynamic texture

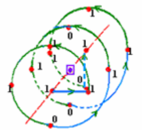
- Dynamic Textures (DT): Temporal texture
- Textures with motion
- An extension of texture to the temporal domain
- Encompass the class of video sequences that exhibit some stationary properties in time

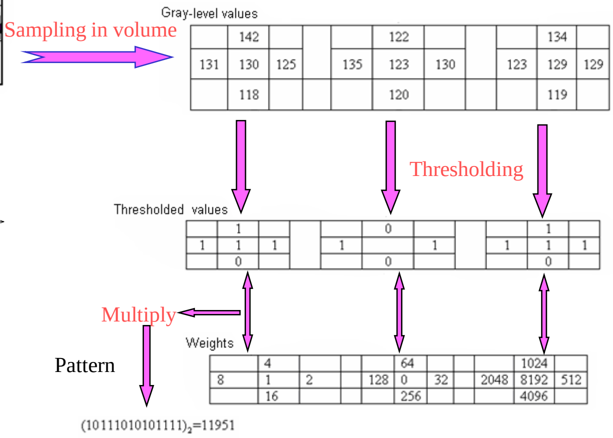
❖ Lots of dynamic textures in real world
❖ Description and recognition of DT is needed

Volume Local Binary Patterns (VLBP)









Gray-level values

	142				122				134	
131	130	125		135	123	130		123	129	129
	118				120				119	

Thresholded values

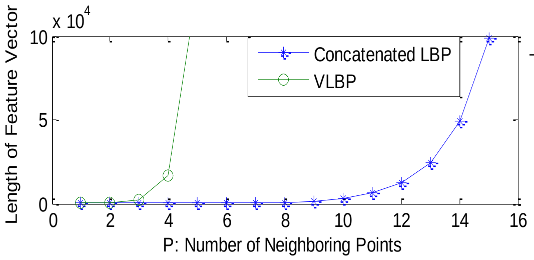
	1				0				1	
1	1	1		1		1		1	1	1
	0				0				0	

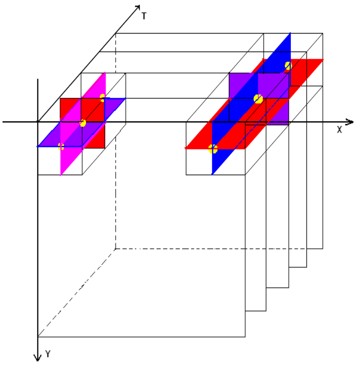
Weights

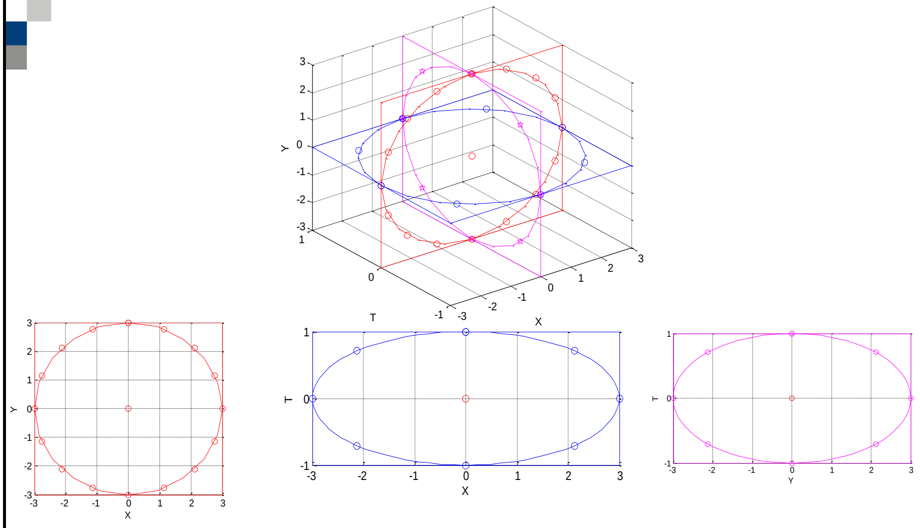
	4				64				1024	
8	1	2		128	0	32		2048	8192	512
	16				256				4096	

$(10111010101111)_2=11951$

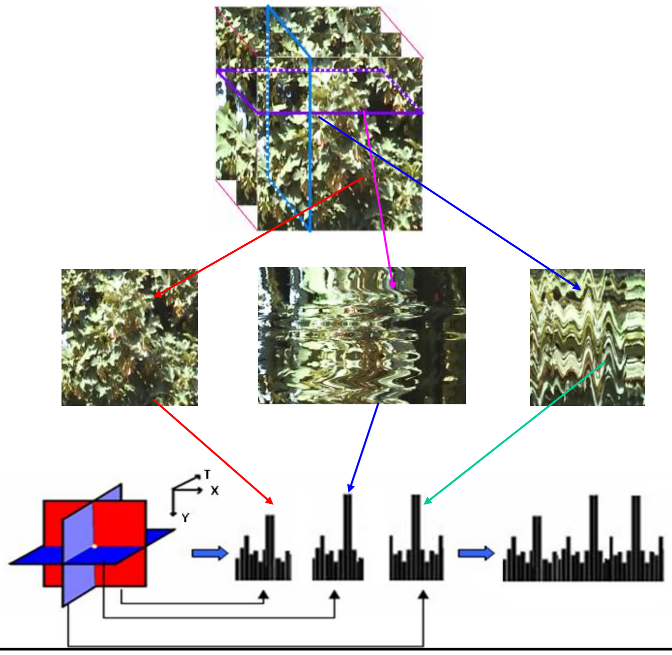
LBP from Three Orthogonal Planes (LBP-TOP)

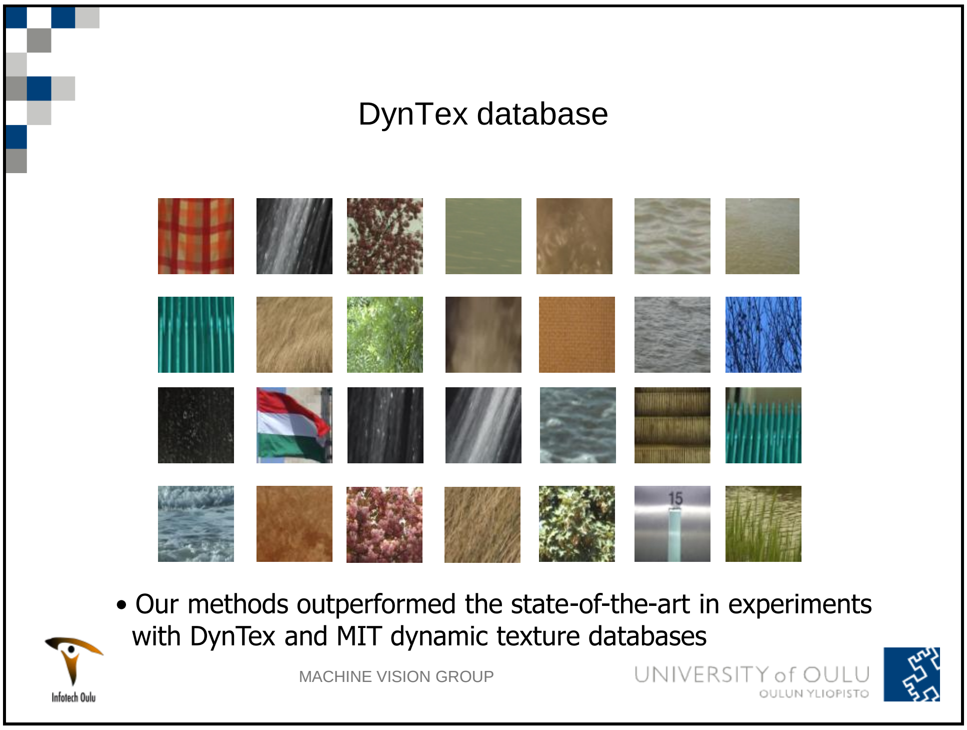






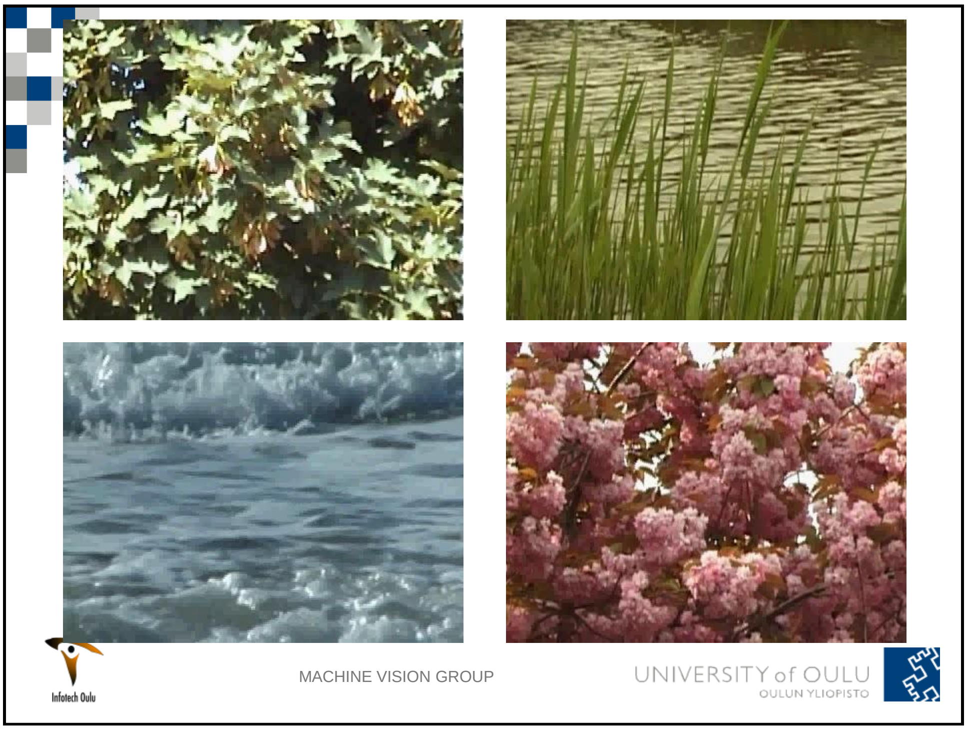
LBP-TOP

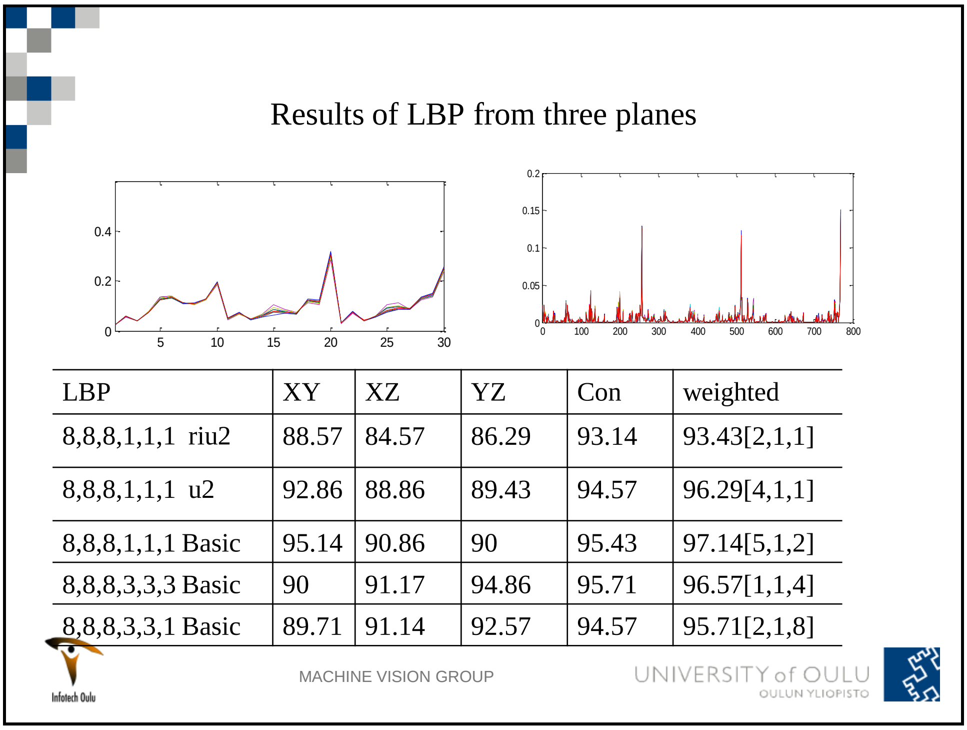




DynTex database

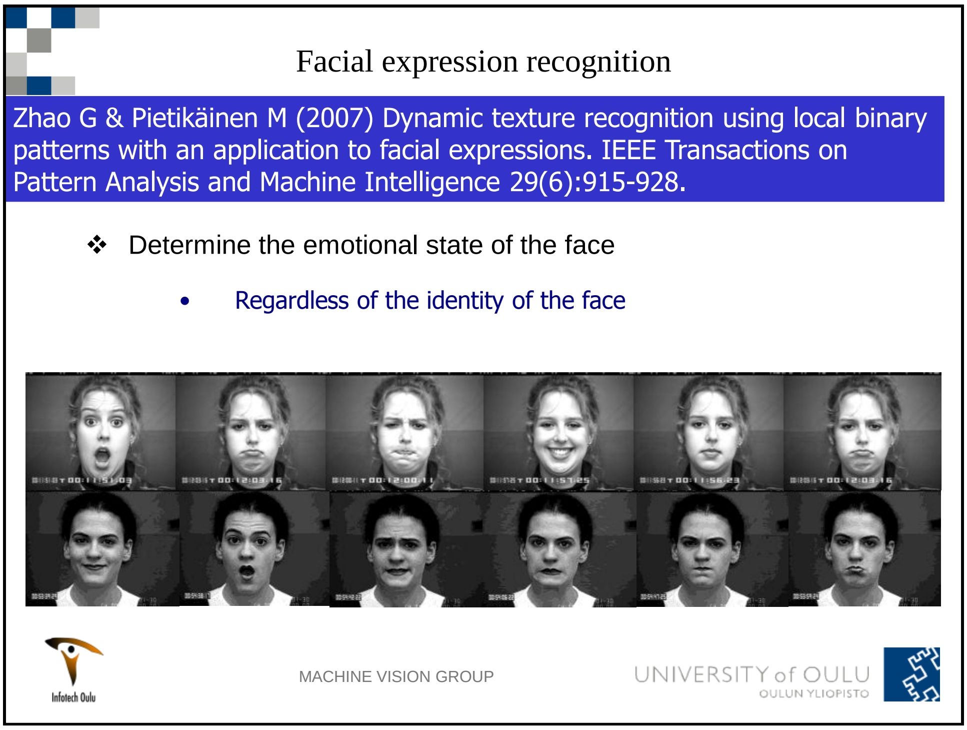
- Our methods outperformed the state-of-the-art in experiments with DynTex and MIT dynamic texture databases





Results of LBP from three planes

LBP	XY	XZ	YZ	Con	weighted
8,8,8,1,1,1 riu2	88.57	84.57	86.29	93.14	93.43[2,1,1]
8,8,8,1,1,1 u2	92.86	88.86	89.43	94.57	96.29[4,1,1]
8,8,8,1,1,1 Basic	95.14	90.86	90	95.43	97.14[5,1,2]
8,8,8,3,3,3 Basic	90	91.17	94.86	95.71	96.57[1,1,4]
8,8,8,3,3,1 Basic	89.71	91.14	92.57	94.57	95.71[2,1,8]

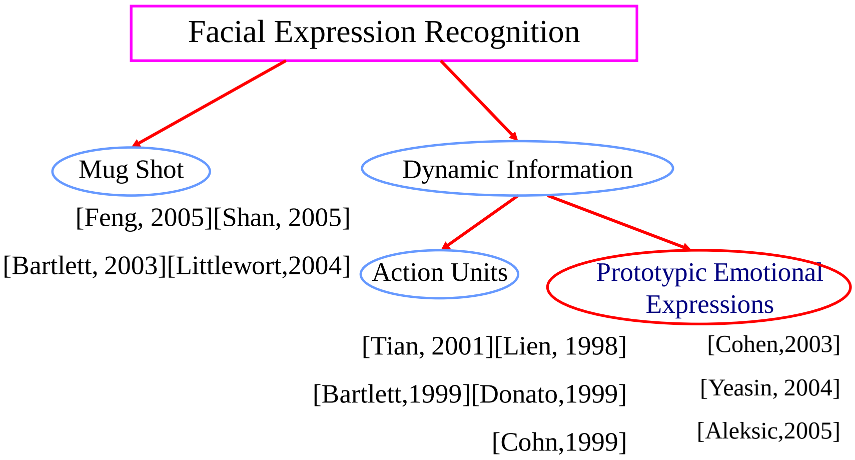


Facial expression recognition

Zhao G & Pietikäinen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6):915-928.

❖ Determine the emotional state of the face

- Regardless of the identity of the face



Facial Expression Recognition

Mug Shot

[Feng, 2005][Shan, 2005]

[Bartlett, 2003][Littlewort,2004]

Dynamic Information

Action Units

[Tian, 2001][Lien, 1998]

[Bartlett,1999][Donato,1999]

[Cohn,1999]

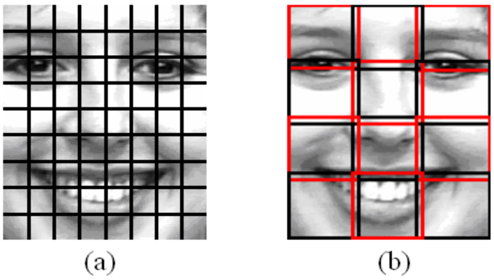
Prototypic Emotional Expressions

[Cohen,2003]

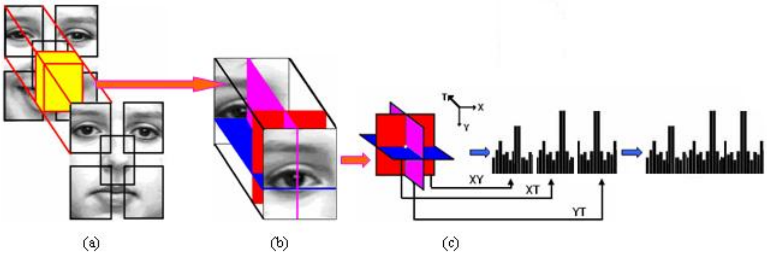
[Yeasin, 2004]

[Aleksic,2005]

Psychological studies [Bassili 1979], have demonstrated that humans do a better job in recognizing expressions from dynamic images as opposed to the mug shot.



(a) Non-overlapping blocks(9 x 8) (b) Overlapping blocks (4 x 3, overlap size = 10)

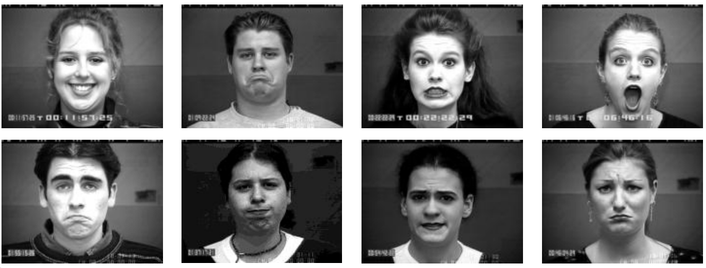


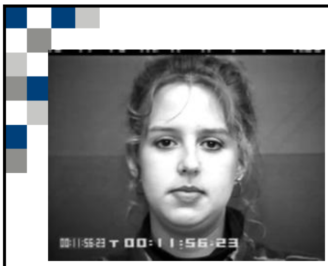
(a) Block volumes (b) LBP features from three orthogonal planes (c) Concatenated features for one block volume with the appearance and motion

Database

Cohn-Kanade database :

- 97 subjects
- 374 sequences
- Age from 18 to 30 years
- Sixty-five percent were female, 15 percent were African-American, and three percent were Asian or Latino.





Happiness

Angry

Disgust

Sadness

Fear

Surprise

Comparison with different approaches

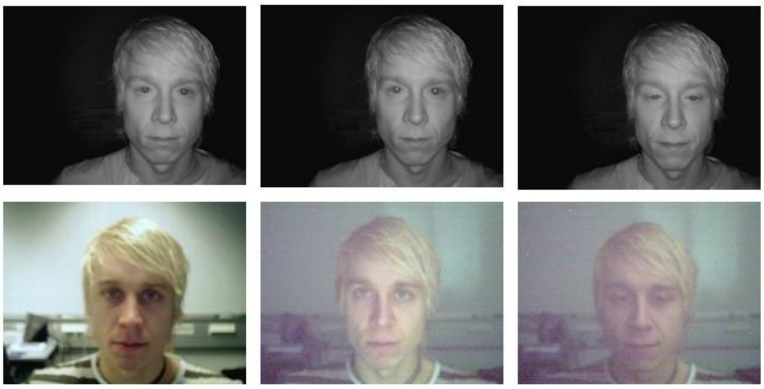
	People Num	Sequence Num	Class Num	Dynamic	Measure	Recognition Rate (%)
[Shan,2005]	96	320	7(6)	N	10 fold	88.4(92.1)
[Bartlett, 2003]	90	313	7	N	10 fold	86.9
[Littlewort, 2004]	90	313	7	N	leave-one-subject-out	93.8
[Tian, 2004]	97	375	6	N	-------	93.8
[Yeasin, 2004]	97	------	6	Y	five fold	90.9
[Cohen, 2003]	90	284	6	Y	-------	93.66
Ours	97	374	6	Y	two fold	**95.19**
Ours	97	374	6	Y	10 fold	**96.26**

Demo for facial expression recognition

- ❖ **Low resolution**
- ❖ **No eye detection**
- ❖ **Translation, in-plane and out-of-plane rotation, scale**
- ❖ **Illumination change**
- ❖ **Robust with respect to errors in face alignment**

Example images in different illuminations

Visible light (VL) : 0.38-0.75 μm
Near Infrared (NIR) : 0.7μm-1.1μm

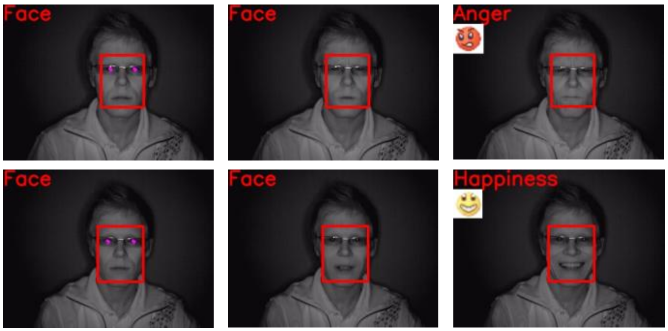


Strong illumination | Weak illumination | Dark illumination

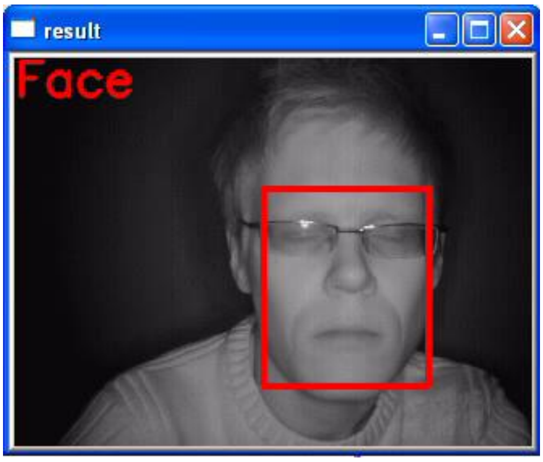
Taini M, Zhao G, Li SZ & Pietikäinen M (2008) Facial expression recognition from near-infrared video sequences. Proc. 19th International Conference on Pattern Recognition (ICPR), 4 p.

On-line facial expression recognition from NIR videos

- NIR web camera allows expression recognition in near darkness.
- Image resolution 320 × 240 pixels.
- 15 frames used for recognition.
- Distance between the camera and subject around one meter.



Start sequences | Middle sequences | End sequences



Facial expression under NIR environment

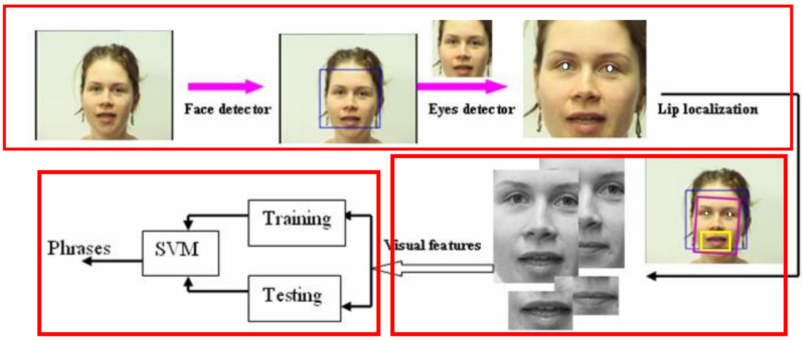
Visual speech recognition

Zhao G, Barnard M & Pietikäinen M (2009). Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia 11(7):1254-1265.

- Visual speech information plays an important role in speech recognition under noisy conditions or for listeners with hearing impairment.
- A human listener can use visual cues, such as lip and tongue movements, to enhance the level of speech understanding.
- The process of using visual modality is often referred to as lipreading which is to make sense of what someone is saying by watching the movement of his lips.

McGurk effect [McGurk and MacDonald 1976] demonstrates that inconsistency between audio and visual information can result in perceptual confusion.

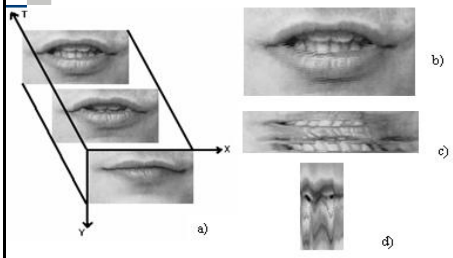
System overview



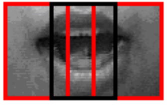
Our system consists of three stages.

- First stage: face and eye detectors, and the localization of mouth.
- Second stage: extracts the visual features.
- Last stage: recognize the input utterance.

Local spatiotemporal descriptors for visual information

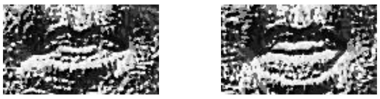


(a) Volume of utterance sequence
(b) Image in XY plane (147x81)
(c) Image in XT plane (147x38) in y =40
(d) Image in TY plane (38x81) in x = 70

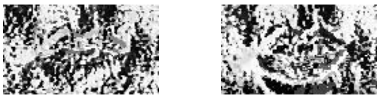


Overlapping blocks (1 x 3, overlap size = 10).

Mouth region images

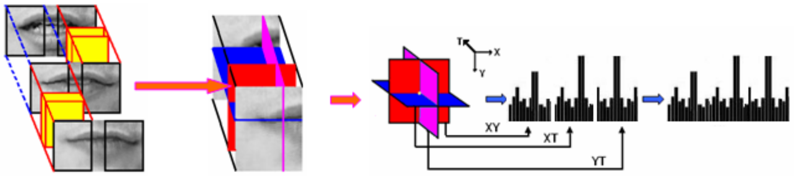


LBP-XY images



LBP-XT images

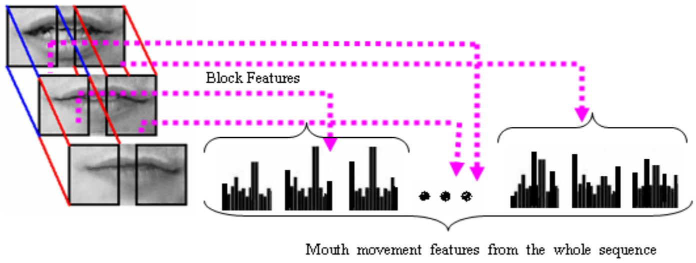
LBP-YT images



(a) (b) (c)

(a) Block volumes appearance and motion (b) LBP features from three orthogonal planes (c) Concatenated features for one block volume with the

Features in each block volume.



Mouth movement representation.

Experiments

- Three databases:

1) Our own visual speech database: OuluVS Database

20 persons; each uttering ten everyday's greetings one to five times.

Totally, 817 sequences from 20 speakers were used in the experiments.

C1	"Excuse me"	C6	"See you"
C2	"Good bye"	C7	"I am sorry"
C3	"Hello"	C8	"Thank you"
C4	"How are you"	C9	"Have a good time"
C5	"Nice to meet you"	C10	"You are welcome"

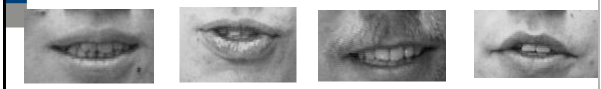
2) Tulips1 audio-visual database

12 subjects, pronouncing the first four digits in English two times in repetition. Totally 96 sequences.

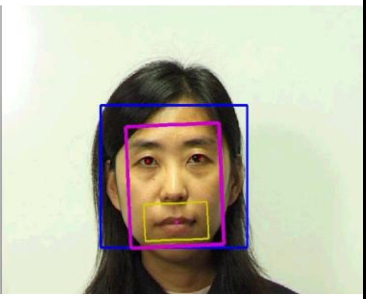
3) AVLetters database

10 people, each uttering 26 english letters three times. Totally 780 sequences.

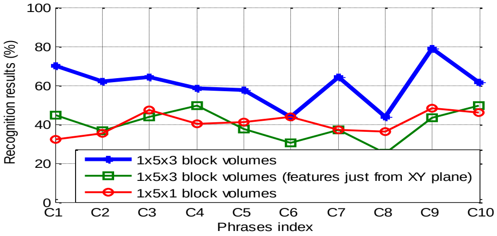
Experimental results - OuluVS database



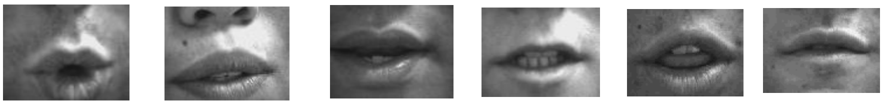
Mouth regions from the dataset.



Speaker-independent:



Experimental results - Tulips1 audio-visual database



Mouth images with translation, scaling and rotation from Tulips1 database.

Comparison to other methods on Tulips1 audio-visual database (speaker independent).

	Features	Normalization	Results (%)
[Arsic 2006]	MRPCA	Y	81.25
[Arsic 2006]	MI MRPCA	Y	87.5
[Gurban 2005]	Temporal Derivatives Features	Y	80 91(a&v, 10 dB SNR level)
Ours	$LBP-TOP_{8,8,8,1,1,1}$ Blocks: 3x6x2	**N**	**92.71**

Visemes	Phonemes	Visemes	Phonemes
/p/	P, B, M	/iy/	IY
/f/	F	/aa/	AA
/t/	T, D, S, Z	/ah/	AY
/ch/	CH, JH, ZH	/ow/	OW
/w/	W, R	/uw/	UW
/k/	K, N, L	/ey/	EH, EY

AVLetters database: 26 letters, 10 people, three utterances per letter.

Confusion matrix from SVMs($LBP-TOP^{u2}_{8,8,8,3,3,3}$ features with 2 × 5 × 3 blocks)

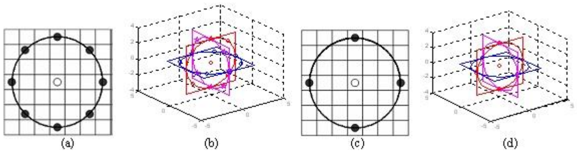
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	17	1			1				1		1	3		5					1							
B	1	18														10						1				
C			11	8	1		1				1					1				5		1				1
D			6	12			1				1									9						1
E	3			1	19						2								2	3						
F						25					1	1	1						1					1		
G			1				16			6										3		2			1	1
H						1		20	1		1	2		1					2		1			1		
I	3					1			18		1	1		2	1			3								
J			1				3			22	1									1		2				
K	2		1	2	1					1	17	1		2						3						
L	4				1			1			1	13	1	5				1		1				2		
M	1					1					1	1	23	1						1						1
N	5			1	2				2			3	1	11		1		1	2					1		
O		1													24		2	1			2					
P	2	10				1								1		15							1			
Q								1							1		17	1			10					
R	1						1		1		1				2			23							1	
S								1						2					19					8		
T			4	4	2						1									19						
U				1					1						3		12				13					
V			3							1	1											23				2
W																					1		28		1	
X						1		1			2			2					6	1				17		
Y		1															1								28	
Z			3	1						2									1	1						22

Principal appearance and motion from boosted spatiotemporal descriptors

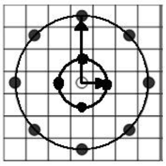
Zhao G & Pietikäinen M (2009) Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. Pattern Recognition Letters 30(12):1117-1127.

Multiresolution features=>Learning for pairs=>Slice selection

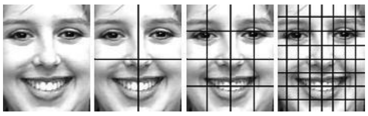
- 1) Use of different number of neighboring points when computing the features in XY, XT and YT slices



- 2) Use of different radii which can catch the occurrences in different space and time scales

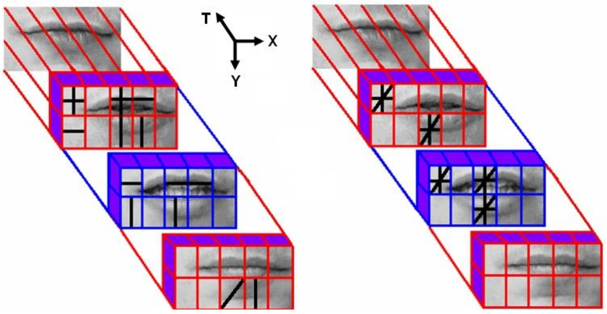


- 3) Use of blocks of different sizes to have global and local statistical features



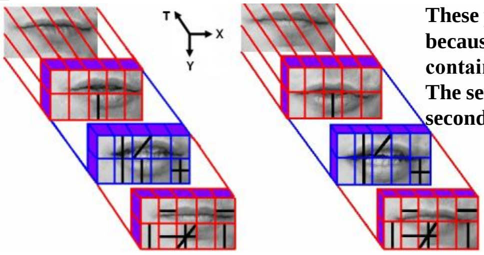
The first two resolutions focus on the
❖ pixel level in feature computation, providing different local spatiotemporal information

the third one focuses on the
❖ block or volume level, giving more global information in space and time dimensions.



Learned first 15 slices (left) and five blocks (right), each block includes three slices from LBP − TOP8,8,8,3,3,3 with 2 × 5 × 3 blocks for all classes learning.

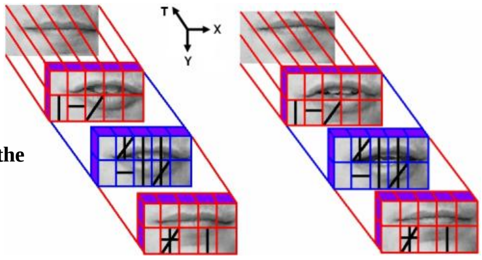
The selected features for all classes are mainly from YT slices (seven out of 15) and XT slices (seven out of 15), just one from XY slices. That suggests that in visual speech recognition the motion information is more important than the appearance.



These phrases were most difficult to recognize because they are quite similar in the latter part containing the same word "you".
The selected slices are mainly in the first and second part of the phrase,

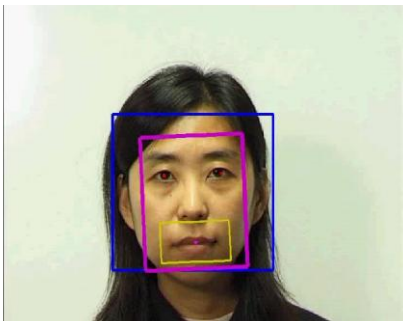
Selected 15 slices for phrases "See you" and "Thank you".

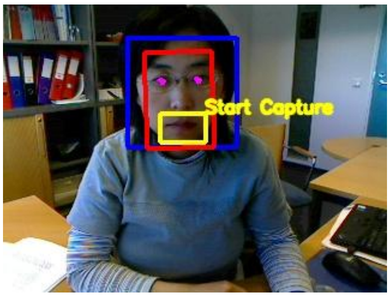
The phrases "excuse me" and "I am sorry"are different throughout the whole utterance, and the selected features also come from the whole pronunciation.



Selected 15 slices for phrases "Excuse me" and "I am sorry".

Demo for visual speech recognition

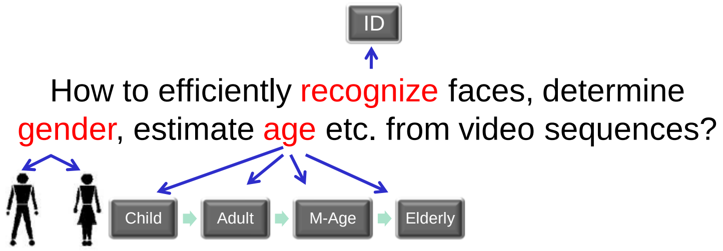




Face recogntion from videos

Hadid A, Pietikäinen M & Li SZ (2007) Learning personal specific facial dynamics for face recognition from videos. In: Analysis of Faces and Gestures, AMGF 2007 Proceedings, Lecture Notes in Computer Science 4778, 1-15.

Problem description



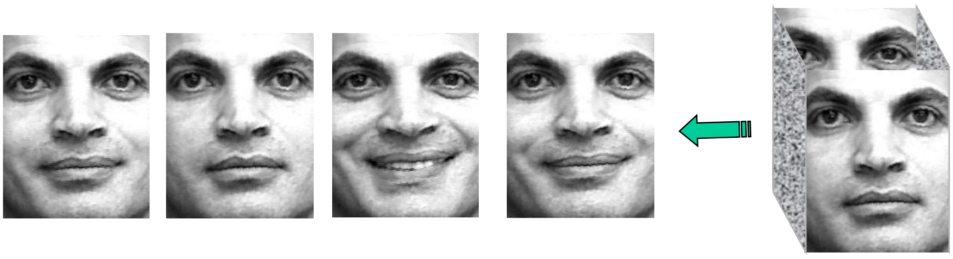
ID

How to efficiently recognize faces, determine gender, estimate age etc. from video sequences?

Child → Adult → M-Age → Elderly

Traditional approaches..

The most common approach is to
apply still image based methods to some selected (or all) frames



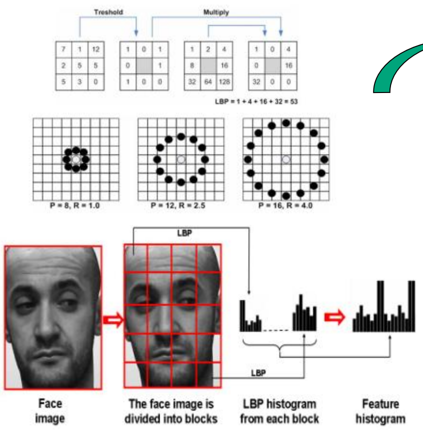
One new direction..

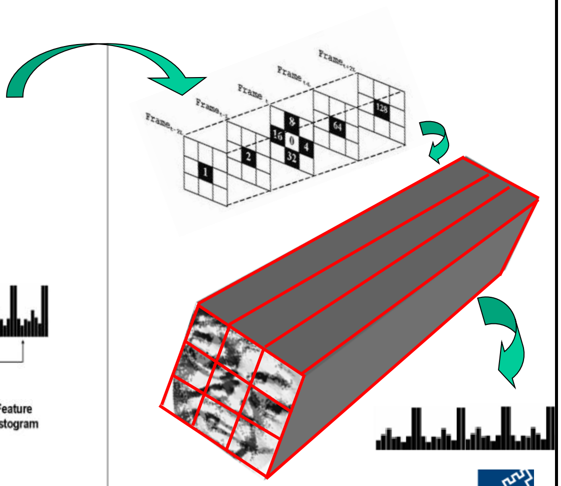
- A Spatiotemporal Approach to Face Analysis from Videos

Motivations:

neuropsychological studies indicating that facial dynamics do support face and gender recognition especially in degraded viewing conditions such as poor illumination, low image resolution…

A face sequence can be seen as a collection of rectangular prisms (volumes) from which we extract local histograms of *Extended* Volume Local Binary Pattern code occurrences.

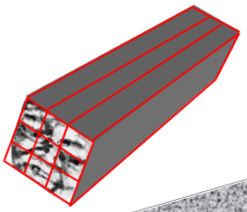


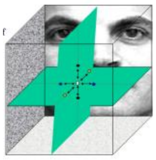


A spatiotemporal approach to face analysis from videos..

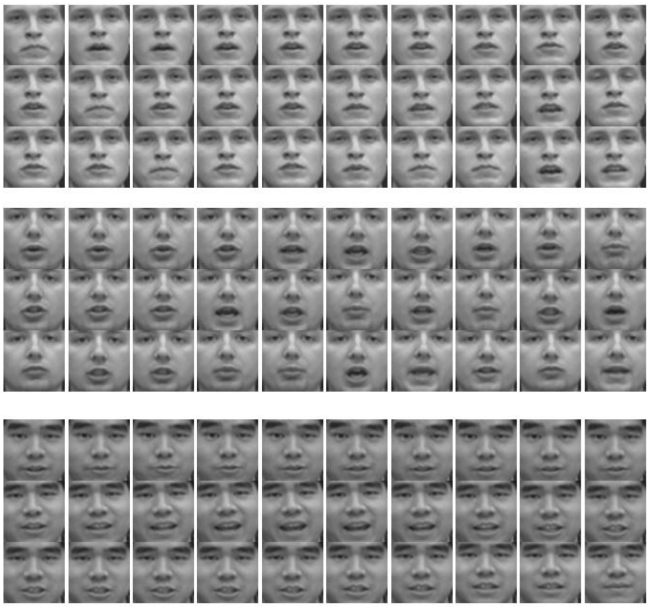
Algorithm:

1. **Divide the video into local prisms**
2. **Consider 3D neighborhood of each pixel**
3. **Apply VLBP**
4. **Feature Selection using AdaBoost**
5. **Extract local histograms**
6. **Histogram concatenation & normalization**
7. **Matching**





Some experimental results



Experiments on face recognition

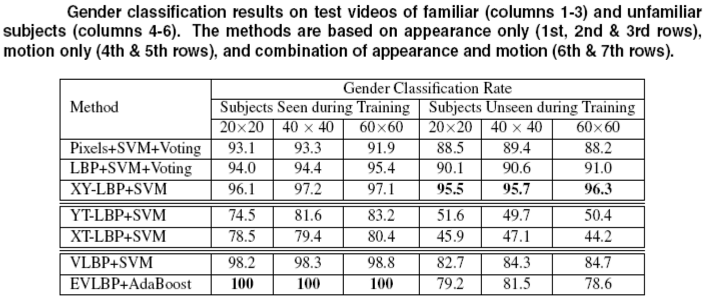
Method	Results on MoBo	Results on Honda/UCSD	Results on CRIM
PCA	87.1%	69.9%	89.7%
LDA	90.8%	74.5%	91.5%
LBP [13]	91.3%	79.6%	93.0%
HMM [8]	92.3%	84.2%	85.4%
ARMA [7]	93.4%	84.9%	80.0%
VLBP [14]	90.3%	78.3%	88.7%
VLBP+AdaBoost	96.5%	89.1%	94.4%
EVLBP+AdaBoost	**97.9%**	**96.0%**	**98.5%**

Static image based versus spatiotemporal based approaches to face recognition

Experiments on gender classification

Hadid A & Pietikäinen M (2009) Combining appearance and motion for face and gender recognition from videos. Pattern Recognition 42:2818-2827.

Databases: CRIM, VidTIMIT and Cohn-Kanade

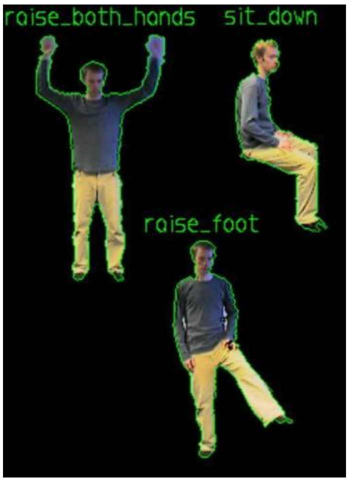


Gender classification results on test videos of familiar (columns 1-3) and unfamiliar subjects (columns 4-6). The methods are based on appearance only (1st, 2nd & 3rd rows), motion only (4th & 5th rows), and combination of appearance and motion (6th & 7th rows).

Method	Gender Classification Rate					
	Subjects Seen during Training			Subjects Unseen during Training		
	20×20	40 × 40	60×60	20×20	40 × 40	60×60
Pixels+SVM+Voting	93.1	93.3	91.9	88.5	89.4	88.2
LBP+SVM+Voting	94.0	94.4	95.4	90.1	90.6	91.0
XY-LBP+SVM	96.1	97.2	97.1	**95.5**	**95.7**	**96.3**
YT-LBP+SVM	74.5	81.6	83.2	51.6	49.7	50.4
XT-LBP+SVM	78.5	79.4	80.4	45.9	47.1	44.2
VLBP+SVM	98.2	98.3	98.8	82.7	84.3	84.7
EVLBP+AdaBoost	**100**	**100**	**100**	79.2	81.5	78.6

Activity recognition

Kellokumpu V, Zhao G & Pietikäinen M (2009) Recognition of human actions using texture. Machine Vision and Applications, in press.

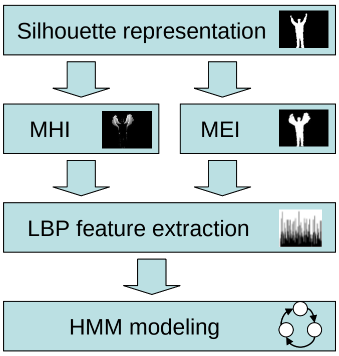


Texture based description of movements

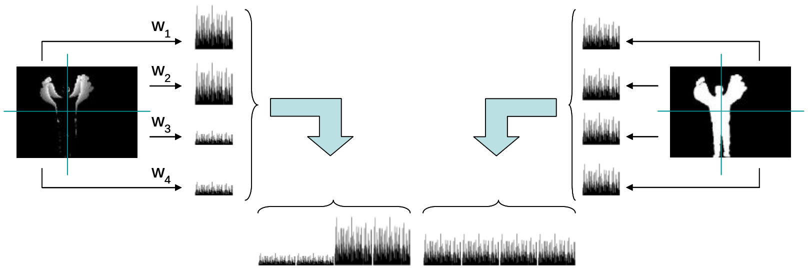
- We want to represent human movement with it’s local properties
  > Texture
- But texture in an image can be anything? (clothing, scene background)
  > Need preprocessing for movement representation
  > We use temporal templates to capture the dynamics
- We propose to extract texture features from temporal templates to obtain a short term motion description of human movement.

Kellokumpu V, Zhao G & Pietikäinen M (2008) Texture based description of movements for activity analysis. Proc. International Conference on Computer Vision Theory and Applications (VISAPP), 1:206-213.

Overview of the approach

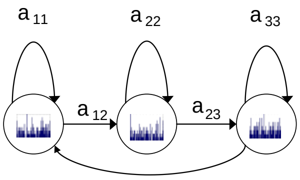


Features



Hidden Markov Models (HMM)

- Model is defined with:
  - Set of observation histograms H
  - Transition matrix A
  - State priors
- Observation probability is taken as intersection of the observation and model histograms:



$$P(h_{obs} \mid s_t = q_i) = \sum \min(h_{obs}, h_i)$$

Experiments

- Experiments on two databases:
  - Database 1:
    - 15 activities performed by 5 persons

  - Database 2 - Weizmann database:
    - 10 Activities performed by 9 persons
    - Walkig, running, jumping, skipping etc.

Experiments – HMM classification

- Database 1 – 15 activities by 5 people
- LBP$_{8,2}$

MHI	99%
MEI	90%
MHI + MEI	100%

- Weizmann database – 10 activities by 9 people
- LBP$_{4,1}$

Ref.	Act.	Seq.	Res.
Our method	**10**	**90**	**97.8%**
Wang and Suter 2007	10	90	**97.8%**
Boiman and Irani 2006	9	81	97.5%
Niebles et al 2007	9	83	72.8%
Ali et al. 2007	9	81	92.6%
Scovanner et al. 2007	10	92	82.6%

Experiments – Continuous data

- Detection and recognition experiments on database 1 using a sliding window based detection.

- Demo

Activity recognition using dynamic textures

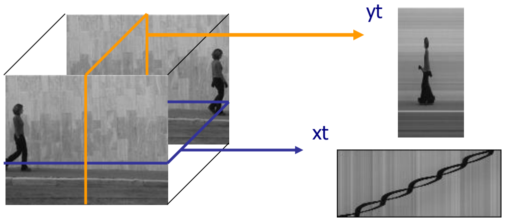
- Instead of using a method like MHI to incorporate time into the description, the dynamic texture features capture the dynamics straight from image data.

- When image data is used, accurate segmentation of the silhouette is not needed
  - Instead a bounding box of a person is sufficient!!

Kellokumpu V, Zhao G & Pietikäinen M (2008) Human activity recognition using a dynamic texture based method. Proc. British Machine Vision Conference (BMVC), 10 p.

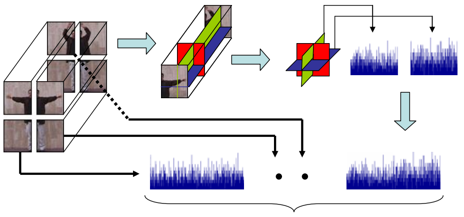
Dynamic textures for action recognition

- Illustration of xyt-volume of a person walking



Dynamic textures for action recognition

- Formation of the feature histogram for an *xyt* volume of short duration



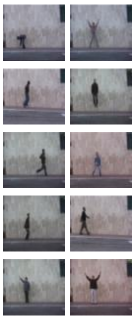
Feature histogram of a bounding volume

- HMM is used for sequential modeling

Action classification results – Weizmann dataset

- Classification accuracy 95.6% using image data

	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	1.00									
Jack		1.00								
Jump			78			.11	.11			
Pjump				1.00						
Run					1.00					
Side						1.00				
Skip					.22		78			
Walk								1.00		
Wave1									1.00	
Wave2										1.00

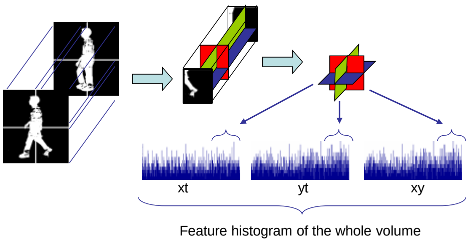


Action classification results - KTH

- Classification accuracy 93.8% using image data

	Box	Clap	Wave	Jog	Run	Walk
Box	.967	.033				
Clap	.003	.987	.01			
Wave	.003	.020	.977			
Jog				.860	,108	.032
Run				.145	.855	
Walk				.020		.980

Dynamic textures for gait recognition



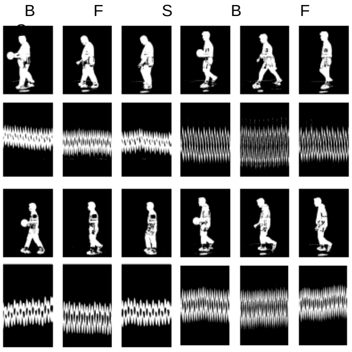
$$Similarity = \sum \min(h_i, h_j)$$

Kellokumpu V, Zhao G & Pietikäinen M (2009) Dynamic texture based gait recognition. In: Advances in Biometrics, ICB 2009 Proceedings, Lecture Notes in Computer Science 5558, 1000-1009.

Experiments - CMU gait database

CMU database

- 25 subjects
- 4 different conditions (ball, slow, fast, ~~incline~~)

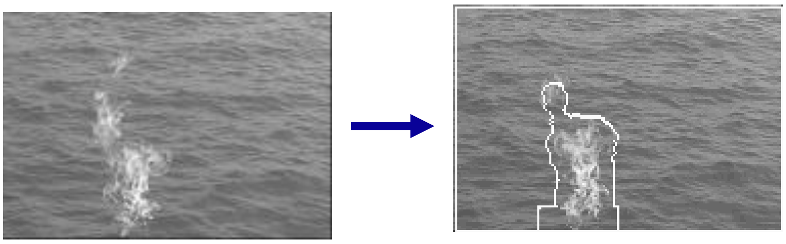


Experiments - Gait recognition results

	S/B	B/S	F/B	B/F	S/F	F/S
CMU [4]	**92 %**	-	-	-	76 %	-
UMD [5]	48 %	68 %	48 %	48 %	80 %	84 %
MIT [6]	50 %	-	-	-	64 %	-
SSP [7]	-	-	-	-	54 %	32 %
SVB frieze [8]	77 %	**89 %**	61 %	73 %	82 %	80 %
LBP-TOP	75 %	83 %	**75 %**	**83 %**	**88 %**	**88 %**

Unsupervised dynamic texture segmentation

Chen J, Zhao G & Pietikäinen M (2008) Unsupervised dynamic texture segmentation using local spatiotemporal descriptors. Proc. International Conference on Pattern Recognition (ICPR), 4 p.



Input

Output

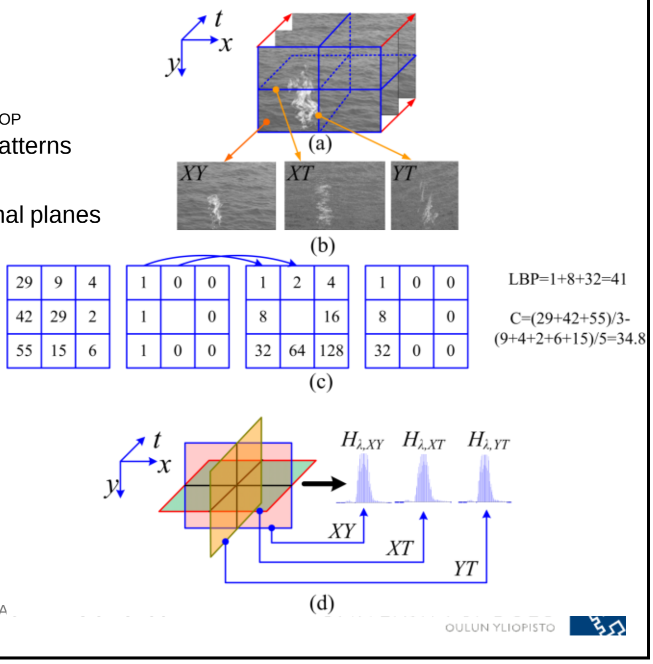
Dynamic texture segmentation

- Potential applications: Remote monitoring and various type of surveillance in challenging environments:
  - monitoring forest fires to prevent natural disasters
  - traffic monitoring
  - homeland security applications
  - animal behavior for scientific studies.

Related work

- Mixtures of dynamic texture model
  - A.B. Chan and N. Vasconcelos, PAMI2008
- Mixture of linear models
  - L. Cooper, J. Liu and K. Huang, Workshop in ICCV2005
- Multi-phase level sets
  - D. Cremers and S. Soatto, IJCV2004
- Gauss-Markov models and level sets
  - G. Doretto, A. Chiuso, Y. N. Wu and S. Soatto, ICCV2003
- Ising descriptors
  - A. Ghoreyshi and R. Vidal, ECCV2006
- Optical flow
  - R. Vidal and A. Ravichandran, CVPR2005

- Feature: (LBP/C)$_{TOP}$
  - Local binary patterns
  - Contrast
  - three orthogonal planes



29	9	4
42	29	2
55	15	6

1	0	0
1		0
1	0	0

1	2	4
8		16
32	64	128

1	0	0
8		0
32	0	0

LBP=1+8+32=41

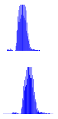
C=(29+42+55)/3-(9+4+2+6+15)/5=34.8

Infotech Oulu

Measure

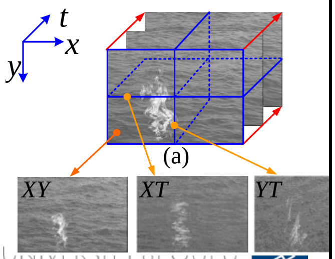
- Similarity measurement

$$\Pi(H_1, H_2) = \sum_{i=1}^{L} \min(H_{1,i}, H_{2,i})$$



- Distance between two sub-blocks

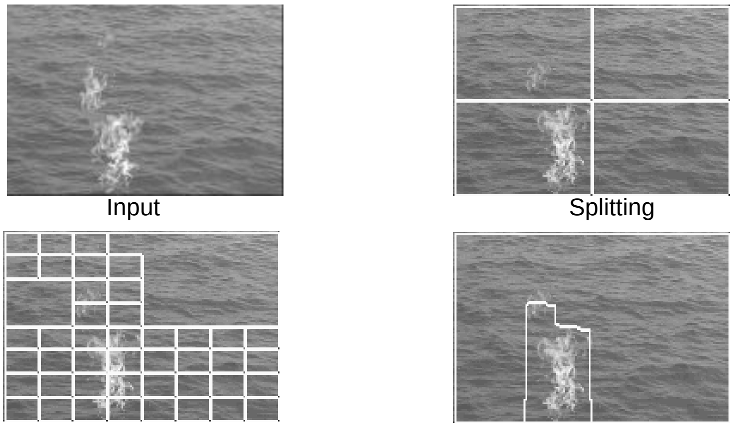
d={$\Pi_{LBP,\ XY}$, $\Pi_{LBP,\ XT}$, $\Pi_{LBP,\ YT}$, $\Pi_{C,\ XY}$, $\Pi_{C,\ XT}$, $\Pi_{C,\ YT}$ }^{T}.



Infotech Oulu

DT segmentation

– Three phases:
  Splitting, Merging, Pixelwise classification.



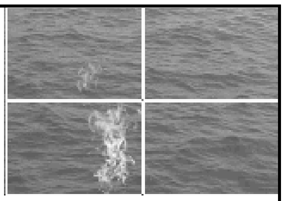
Input

Splitting

Merging

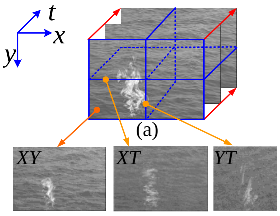
Pixelwise classification

Splitting

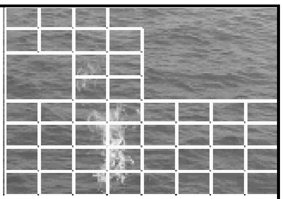


- Recursively split each input frame into square blocks of varying size.

- criterion of splitting:
  - *one* of the features in the three planes (i.e., LBPπ and Cπ, π=*XY*, *XT*, *YT*) votes for splitting of current block

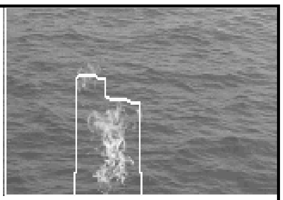


Merging



- Merge those similar adjacent regions with smallest merger importance (*MI*) value
- $MI : MI = f(p) \times (1-\Pi)$
  - Π is the distance between two regions
  - $f(p)$= sigmoid(βp). (β=1, 2, 3, …)
    - $p = N_b / N_f$
    - N_b is the number of pixels in current block
    - N_f is the number of pixels in current frame

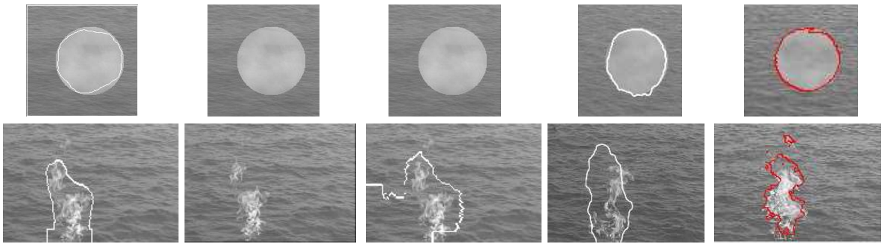
Pixelwise classification



- Compute $(LBP/C)_{TOP}$ histograms over its circular neighbor for each boundary pixel.
- Compute the similarity between neighbors and connected models.
- Re-label the pixel if the label of the nearest model votes a different label.

Experimental results

Some results on types of sequences and compared with existing methods.

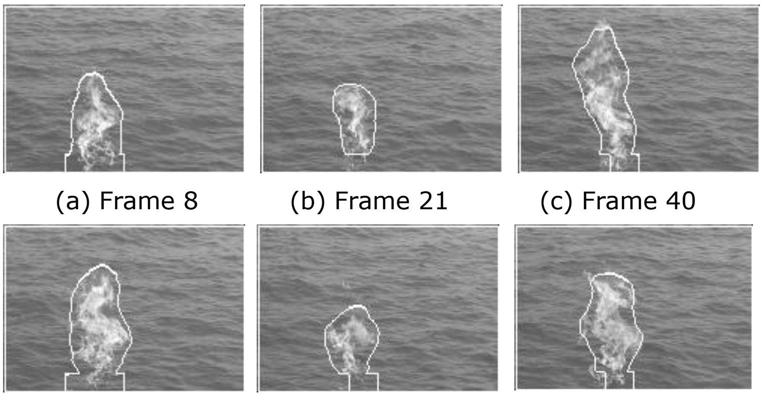


(a) Our method (b) LBP/C (c) LBP-TOP (d) Method in [6] (e) Method in [7]

[6] G. Doretto, A. Chiuso, Y. N. Wu and S. Soatto, Dynamic Texture Segmentation, *ICCV*, 2003
[7] A. Ghoreyshi and R. Vidal, Segmenting Dynamic Textures with Ising Descriptors, ARX Models and Level Sets, *ECCV*, 2006

Experimental results

- Results on sequences *ocean-fire-small*



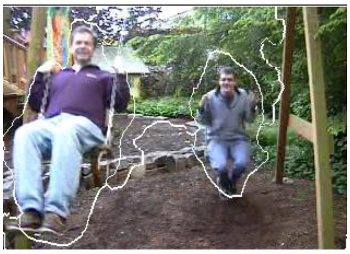
(a) Frame 8 (b) Frame 21 (c) Frame 40

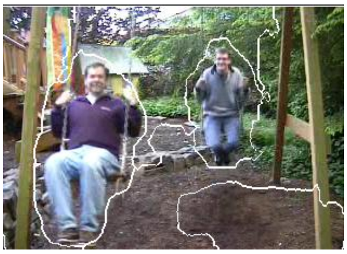
(d) Frame 60 (e) Frame 80 (f) Frame 100

Experimental results

Chen J, Zhao G & Pietikäinen M (2009) An improved local descriptor and threshold learning for unsupervised dynamic texture segmentation. Proc. ICCV Workshop on Machine Learning for Vision-based Motion Analysis, 460-467.

- Results on a real challenging sequence


(a) Frame 5


(b) Frame 10

Dynamic texture synthesis

Guo Y, Zhao G, Chen J, Pietikäinen M & Xu Z (2009) Dynamic texture synthesis using a spatial temporal descriptor. Proc. IEEE International Conference on Image Processing (ICIP), 2277-2280.

- Dynamic texture synthesis is to provide a continuous and infinitely varying stream of images by doing operations on dynamic textures.

Introduction

- **Basic approaches to synthesize dynamic textures:**

  - parametric approaches

- physics-based
- method and image-based method

  - nonparametric approaches: they copy images chosen from original sequences and depends less on texture properties than parametric approaches

- **Dynamic texture synthesis has extensive applications in:**

  - video games
  - movie stunt
  - virtual reality

Synthesis of dynamic textures using a new representation

A. Schödl, R. Szeliski, D. Salesin, and I. Essa, “Video textures,” in Proc. ACM SIGGRAPH, pp. 489-498, 2000.

- The basic idea is to create transitions from frame i to frame j anytime the successor of i is similar to j, that is, whenever $D_{i+1,\ j}$ is small.

- **The algorithm of the dynamic texture synthesis:**

1. **Frame representation;**
2. **Similarity measure;**
3. Distance mapping;
4. Preserving dynamics;
5. Avoid dead ends;
6. Synthesis

Calculate the concatenated local binary pattern histograms from three orthogonal planes for each frame of the input video

Compute the similarity measure Dij between frame pair *Ii* and *I j* by applying Chi-square to the histogram of representation

To create transitions from frame *i* to *j* when *i* is similar to *j* , all these distances are mapped to probabilities through an exponential function Pij. The next frame to display after *i* is selected according to the distribution of *Pij*.

Match subsequences by filtering the difference matrix Dij with a diagonal kernel with weights $[w-m,...,wm-1]$

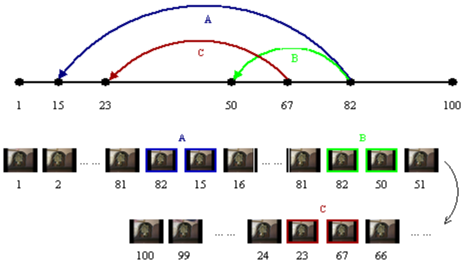
Distance measure can be updated by summing future anticipated costs

When transitions of video texture are identified, video frames are played by video loops

Synthesis of dynamic textures using a new representation

An example:

Considering that there are three transitions: $i_n \rightarrow j_n$ ($n = 1 , 2 , 3$) , loops from the source frame *i* to the destination frame *j* would create new image paths, named as loops. A created cycle is shown as:



Transitions	(i_n, j_n)
n=1	(82, 15)
n=2	(82, 50)
n=3	(67, 23)

Experiments

- We have tested a set of dynamic textures, including natural scenes and human motions.

  (http://www.texturesynthesis.com/links.htm and DynTex database, which provides dynamic texture samples for learning and synthesizing.)

- The experimental results demonstrate our method is able to describe the DT frames from not only space but also time domain, thus can reduce discontinuities in synthesis. (http://www.ee.oulu.fi/~guoyimo/download/)

Experiments

- Dynamic texture synthesis of natural scenes concerns temporal changes in pixel intensities, while **human motion synthesis** concerns temporal changes of body parts.

- The synthesized sequence by our method maintains **smooth dynamic behaviors**. The good performance demonstrates its ability to synthesize complex human motions.

Summary

- Modern texture operators form a generic tool for computer vision
- LBP and its spatiotemporal extensions are very effective for various tasks in computer vision
- Spatiotemporal LBP descriptors combine appearance and motion

- The advantages of the LBP methods include
  - computationally very simple
  - can be easily tailored to different types of problems
  - robust to illumination variations
  - robust to localization errors

- For a bibliography of LBP-related research, see
  http://www.ee.oulu.fi/research/imag/texture

Example applications using or inspired by spatiotemporal LBPs

- Recognition of dynamic textures: Zhao & Pietikäinen, PAMI 2007
- Segmentation of dynamic textures: Chen et al., ICPR 2008, MLVMA 2009
- Facial expression recognition: Zhao & Pietikäinen, PAMI 2007, PRL 2009; Yang et al., PRL 2009
- Face and gender recognition: Hadid & Pietikäinen, AMFG 2007, PR 2009
- Visual speech recognition: Zhao et al., IEEE T Multimedia 2009
- Analysis of facial paralysis: He et al., IEEE T Biomed. Eng. 2009
- Backgroud subtraction: Zhong et al., JCIS 2008
- Recognition of actions: Kellokumpu et al., BMVC 2008, MVA 2009
- Recognition of events: Ma & Cisar, ViSU 2009
- Recognition of actions using a sparse descriptor: Mattivi & Shao, CAIP 2009
- Gait recognition: Kellokumpu et al., ICB 2009
- Driver fatigue detection: Yin et al., IJPRAI 2009
- Video texture synthesis: Guo et al., ICIP 2009

Thanks!