Event modeling and recognition in video

N. Gkalelis

Outline

- Problem formulation
- Joint content-event model

- Referencing mechanism
  - discriminant analysis + model vectors
  - Experimental results
    - Event-based shot classification
    - Entity (face) classification
    - Event-based video classification

- Conclusions

Problem formulation

- The amount of digital video content has grown in unprecedented levels
- Effective exploitation of this information is very important in many application domains: surveillance, entertainment, World Wide Web, and other
- QoS of current video processing tools is far beyond the required satisfaction levels

- To increase QoS we need:
  - Human friendly representation models
  - Effective video analysis algorithms for model enrichment:
    - Efficient manipulation of large scale video collections
    - Accurate responses to user request

Why focus on events ?

- Effective management of video content is hindered by the semantic gap:
  - Difference between what a human understands by e.g. viewing an image and what meaning a machine can automatically extract from it

- One important aspect of the semantic gap relates to events
  - Machines consider low-level features, objects, elementary actions,…
  - People structure their memory mostly based on *high-level events*

➢ *Event-based models may narrow the semantic gap*

Model aspects

- ✓ What aspects should be covered by an event model [1] ?

- Formality: Formal definition of concepts and relations, e.g., using a foundational ontology

- Informational aspect: Information regarding the event itself, e.g., name, type, etc.

[1] U. Westermann and R. Jain, “Toward a common event model for multimedia applications”, IEEE Multimedia, vol. 14, no. 1, pp. 19–29, Jan. 2007.

Model aspects

- Experiential aspect: Multimedia data comprising the experiential dimension of the event.
- Temporal aspect: Relative or absolute time regarding the occurrence of the event
- Spatial aspect: Absolute or relative location of the event
- Compositional aspect: Composite events made up of other events.

Model aspects

- Causal aspect: A change on an event may trigger the state of another event
- Interpretation aspect: The same event perceived differently by different people (depending on their past experience and the given situation)
- Uncertainty aspect: Uncertain results of automatic event indexing algorithms (hardly produce results with 100% confidence)
- Complexity aspect: Requirement of low complexity, and easily processable event descriptions by using the event model

Related work

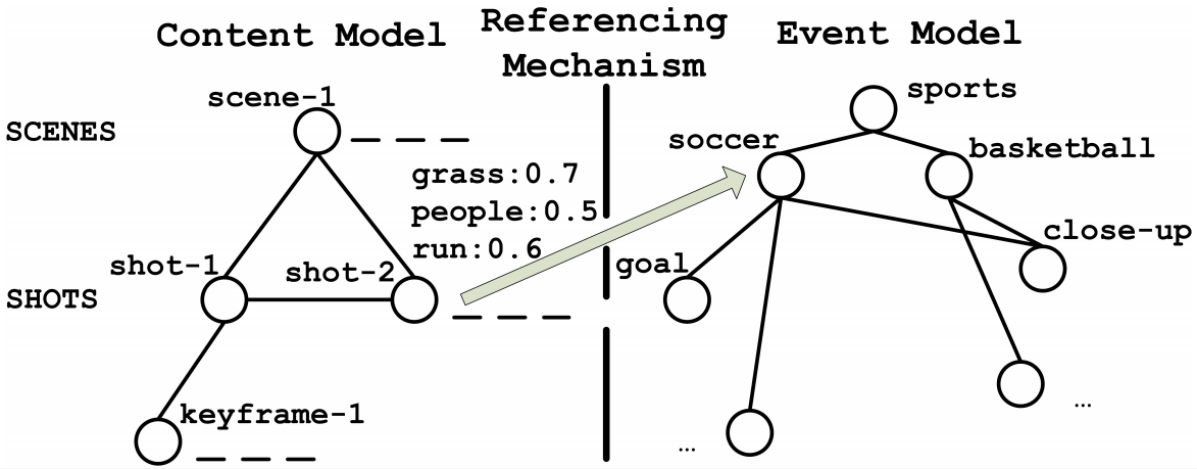
		[11]	[12]	[3]	[4]	[5]	[6]	[28]	[9]	[29]	[30]
Formality aspect		Lim.	No	No	No	No	No	Yes	Yes	No	No
Informational aspect		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Experiential aspect	**Media decomposition**	Yes	Yes	Lim.	Lim.	–	No	–	–	–	–
	Media independence	No	No	Lim.	Lim.	–	Yes	–	–	–	–
Temporal aspect	**Absolute**	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Relative	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes
Spatial aspect	**Absolute**	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
	Relative	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Compositional aspect		Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Casual aspect		Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes
Interpretation aspect		No	No	Yes	No	No	Yes	Yes	No	No	No
Uncertainty aspect		No	No	No	Yes	No	Yes	No	No	No	No
Complexity		Avg.	High	Avg.	High	Low	Avg.	High	Low	Low	Low

- Most event models in the literature provide little or no support for capturing the structure of multimedia content
- A few event models have been proposed for annotating video data, however, they treat events as second class entities

✓ To address these limitations, design a joint content-event model for automatic event-based indexing of multimedia content

Joint content-event model

- Content part to describe decomposition of multimedia data
  - Hierarchical graph structure
  - Content nodes represent content segments, e.g., shot, scene, etc.
  - Edges denote temporal or compositional relationships
- Event part to describe real-life event
  - More general graph structure
  - Event nodes represent real-life event elements (participants, sub-events, etc.)
  - Edges denote a variety of relationships (spatial, casual, etc.)

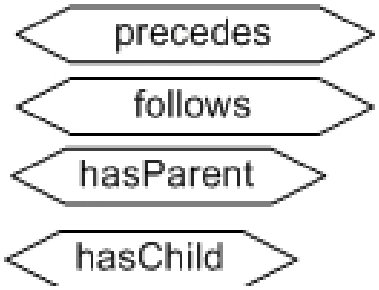


Content node properties

URI indexing content nodes in uniform manner

Content type taxonomy similar to MPEG-7 MDS [1]
(audio segment, speaker segment, video segment, moving region, still region, scene, shot, etc.)

Relative temporal and compositional position of content node in the graph

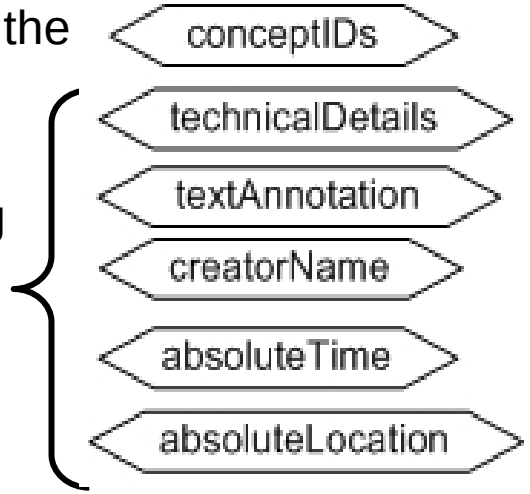


[1] P. Salembier and J. R. Smith. MPEG-7 Multimedia description schemes. IEEE Transactions on Circuits and Systems for Video Technology, 11(6):748-759, Jun 2001.

Content node properties

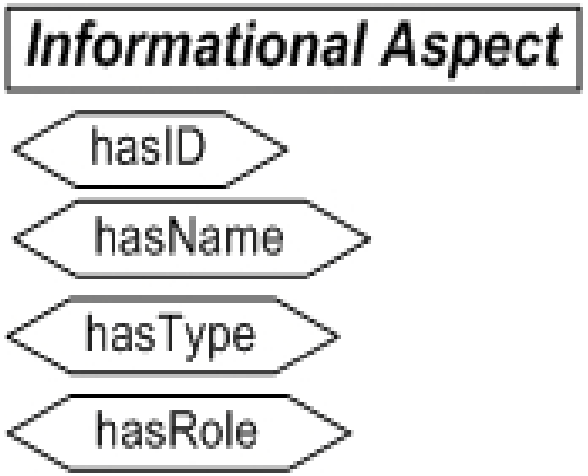
Extracted by analyzing the visual information of the content segment, e.g., grass, run

Information extracted from metadata accompanying the content segment, e.g., frame rate, text annotations, creator name, timestamp, geospatial data, etc.



Actual spatiotemporal position of content segment, e.g., frame range, bounding box

Event node properties



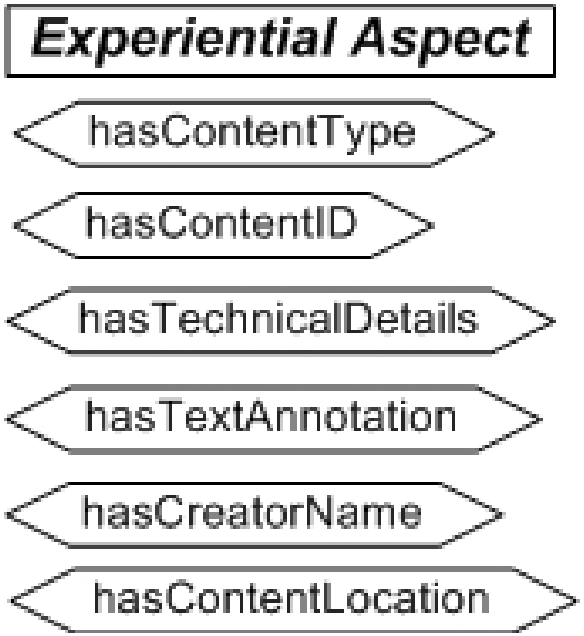
URI to represent event node in global scope

DOLCE types *event*, *agent* and *place* are borrowed to characterize an event element

Role of the event element in a specific context (different roles regarding the situation, e.g., policeman or gunshot victim)

Event node properties

Filled with information directly transferred from respective properties of content node, as soon as the connection between content node and event node is established

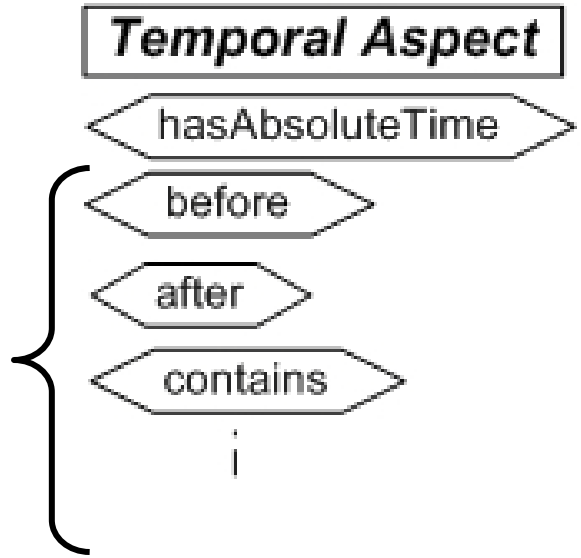


The actual location of the content segment is directly recorded in the content location property

Event node properties

Absolute time of event occurrence is captured using the ISO 8601 standard [1]
1994 - 11 - 05T13:15:30Z
YYYY-MM-DDThh:mm:ss:TZD

Relative time is captured using Allen's Time Calculus relationships (before, meets, overlaps, contains, starts, finishes, equals) [2]



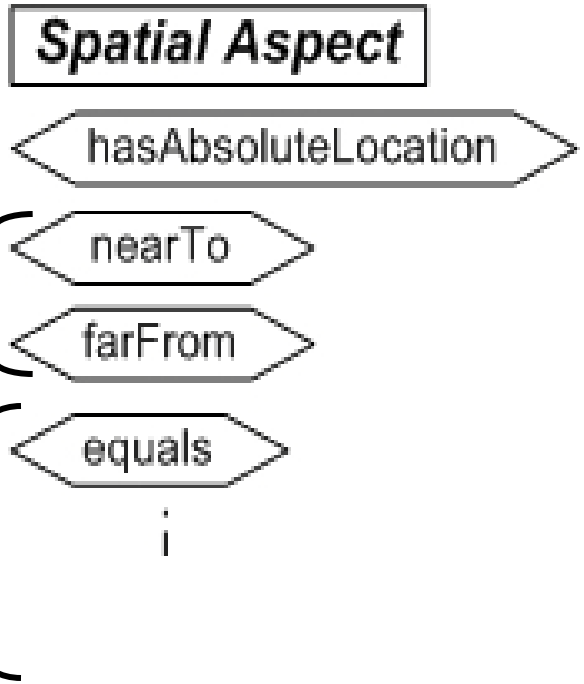
[1] "Date and time formats," http://www.w3.org/TR/NOTE-datetime.
[2] J. F. Allen, "Maintaining knowledge about temporal intervals," *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, Nov. 1983.

Event node properties

Captured in latitude longitude form as defined in Basic Geo Vocablary [1]
<geo:lat>55.701</geo:lat>
<geo:long>12.552</geo:long>

DOLCE properties to capture relative spatial relations

Region Connection Calculus (RCC) to capture more complex relative spatial relations [2]
(equal, disconnection, external connection, partial overlap, tangential proper part (TPP) , non-TPP)



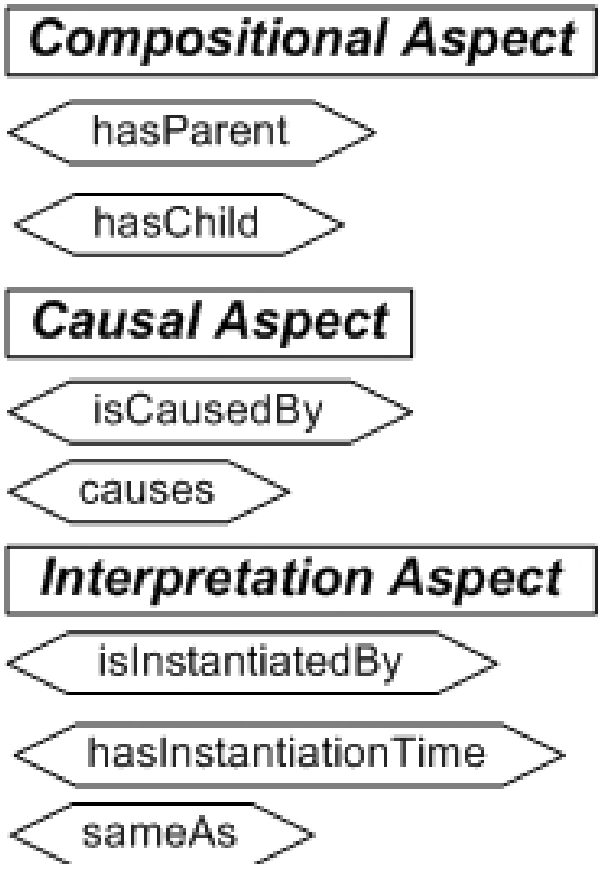
[1] “Basic geo (wgs84 lat/long) vocabulary,” http://www.w3.org/2003/01/geo/
[2] D. A. Randell, Z. Cui, and A. G. Cohn, “A spatial logic based on regions and connection,” in Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning, Jan. 1992, pp. 165–176

Event node properties

Parent and child relationship to indicate composite and component events

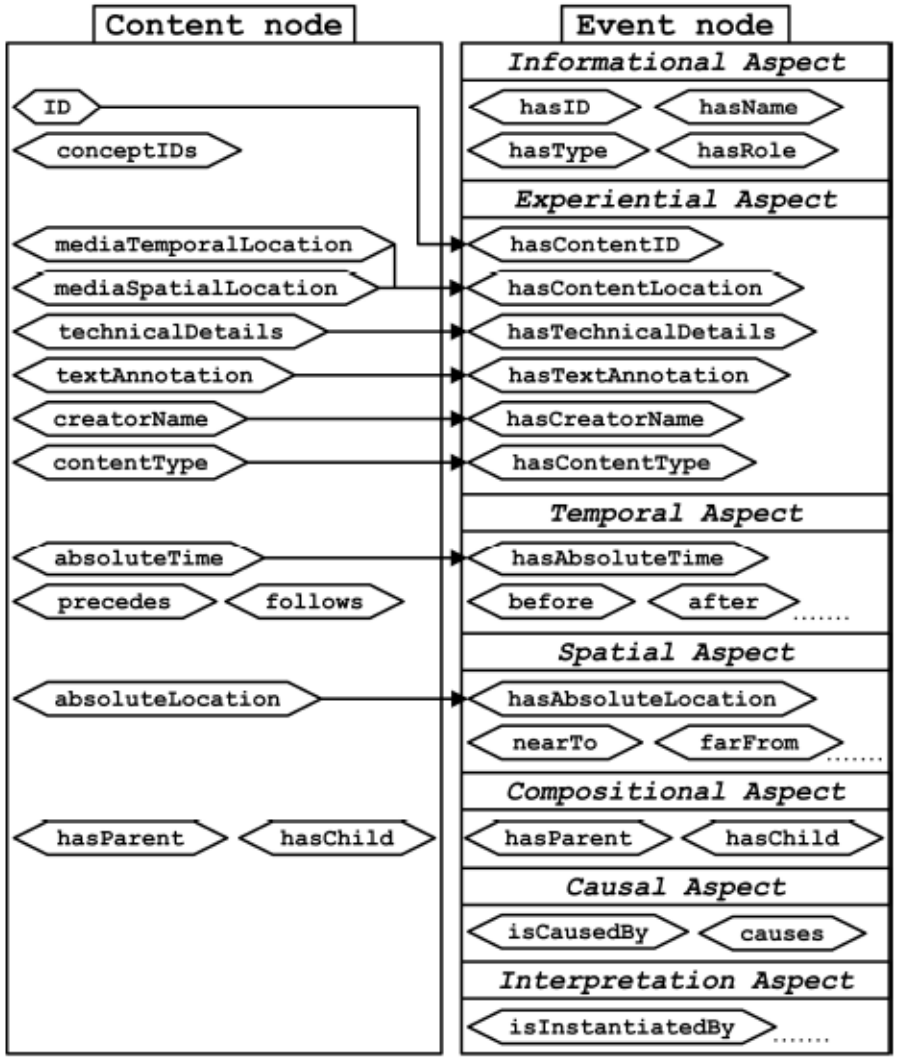
Properties to capture cause-effect relationships, e.g., a robbery event may cause a gunshot event

Functionality to allow users to provide a different interpretation of the same event



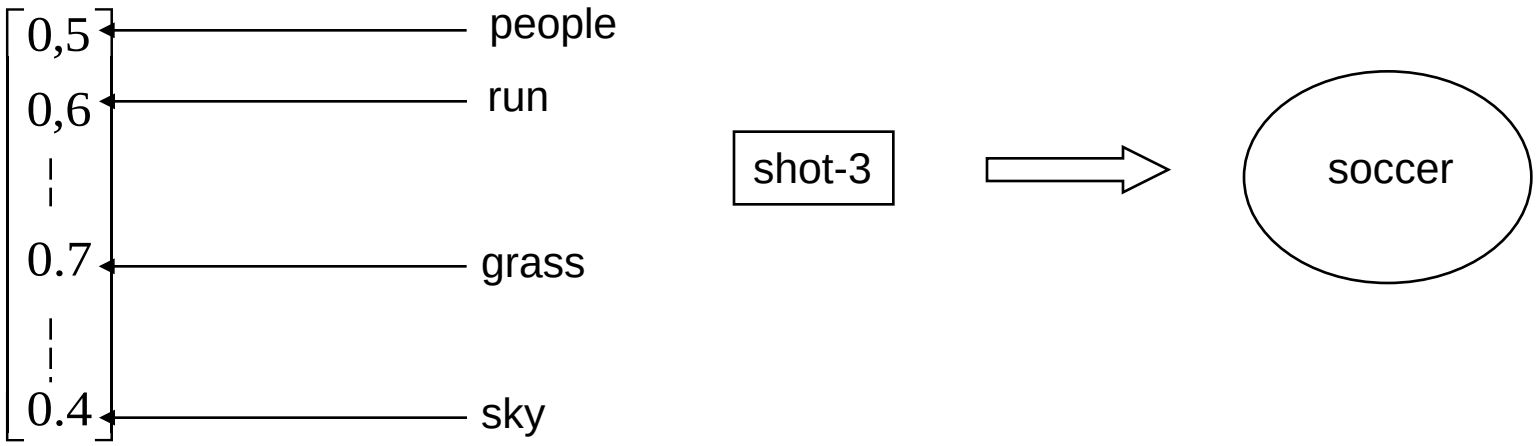
Referencing mechanism

- Exploitation of video analysis algorithms (**visual-model-vector + discriminant-analysis approach**) to automatically connect content nodes with event nodes
- Common properties of content and event nodes to transfer information from one side to other when a connection is established



Example: Establishment of the connection

- A trained referencing mechanism is used to associate content segments with event elements
  - A content segment is represented by a visual model vector using trained concept detectors
  - Model vector values show the DoC that a concept is present in the content segment
  - The referencing mechanism classifies model vectors to event elements



Outline

- Problem formulation
- Joint content-event model

- Referencing mechanism
  - Discriminant analysis + model vectors
  - Experimental results
    - Event-based shot classification
    - Entity (face) recognition
    - Event-based video classification

- Conclusions

Discriminant analysis

- Training set of *N* labeled data belonging to *C* classes is used

$$\mathcal{U} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \qquad \{\mathcal{X}_1, \ldots, \mathcal{X}_C\}$$

- Identify the projection matrix that maps the *F*-dimensional signal to a *D*-dimensional discriminant subspace

$$\mathbf{z}_i = \tilde{\mathbf{\Psi}}^T \mathbf{x}_i \qquad D << F$$

- The transformation matrix is identified by maximizing the following objective function or equivalently solving the generalized eigenvalue problem

$$J(\mathbf{\Psi}) = \frac{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{A} \mathbf{\Psi})}{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{B} \mathbf{\Psi})} \qquad \mathbf{A}\mathbf{\Psi} = \mathbf{B}\mathbf{\Psi}\mathbf{\Lambda}$$

Linear discriminant analysis

- Seeks projection directions that preserve class separability

$$J_{lda}(\mathbf{\Psi}) = \frac{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_b \mathbf{\Psi})}{\mathrm{tr}(\mathbf{\Psi}^T \mathbf{S}_w \mathbf{\Psi})}$$

- Metric to measure distance between different classes

$$\mathbf{S}_b = \sum_{i=1}^{C} p_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

- Metric to measure distance of samples within the same class

$$\mathbf{S}_w = \sum_{i=1}^{C} p_i \mathbf{\Sigma}_i \qquad \mathbf{\Sigma}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in \mathcal{X}_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T$$

Heteroscedastic class distributions

- Fundamental assumption of LDA is that class distributions are homoscedastic
- Kernel extensions of LDA (KLDA) map data in a higher dimensional space where classes are expected to be homoscedastic
- Discover a subclass homoscedastic structure of the data by clustering (less computational expensive)

Subclass discriminant analysis

- Uses a cross-validation (CV) procedure to obtain the *optimal partition of the data*:

  For *r*= 1 to R
  - Partition each class of the training set to *r* subclasses
  - Compute projection matrix that maximizes the criterion

$$J(\mathbf{\Psi})_{msda} = \frac{|\ \mathbf{\Psi}^T \mathbf{S}_{bsb} \mathbf{\Psi}\ |}{|\ \mathbf{\Psi}^T \mathbf{\Sigma}_{\mathrm{X}} \mathbf{\Psi}\ |}$$

  - Between subclass scatter matrix emphasizes separation of subclasses belonging to different classes

$$\mathbf{S}_{bsb} = \sum_{i=1}^{C-1} \sum_{j=1}^{H_i} \sum_{k=i+1}^{C} \sum_{l=1}^{H_k} p_{i,j} p_{k,l} (\boldsymbol{\mu}_{i,j} - \boldsymbol{\mu}_{k,l})(\boldsymbol{\mu}_{i,j} - \boldsymbol{\mu}_{k,l})^T$$

  - Classify samples of validation set and retain CCR and projection matrix

✓ Select optimal partition (projection matrix) as the one that provide maximum CCR

Improved SDA

- Drawback of SDA:
  - At every iteration the total number of subclasses is increased by C

$$H^{(r)} = \sum_{i=1}^{C} H_i^{(r)} = H^{(r-1)} + C$$

  - SDA criterion may be maximized in intermediate number of subclasses
- ISDA
  - Prior to the cross validation approach
    - Sort training samples using a NN-based algorithm

$$\{\mathbf{x}_1^1, \ldots, \mathbf{x}_{N_1}^1, \ldots, \mathbf{x}_\nu^i, \ldots, \mathbf{x}_1^C, \ldots, \mathbf{x}_{N_C}^C\}$$

    - Compute the distance between neighboring vectors

$$\mathbf{d} = [d_1^1, \ldots, d_{N_1}^1, \ldots, d_\nu^i, \ldots, d_1^C, \ldots, d_{N_C-1}^C]^T$$

  - At each CV iteration increase total number of subclasses only by one
    - Use a peak picking algorithm to obtain the positions of the *h* largest peaks

$$\{r_1, \ldots, r_h\} = g(\mathbf{d}, h)$$

    - Use these positions to partition the classes so that

$$H^{(r)} = H^{(r-1)} + 1$$

Event-based shot classification

- MediaMill Challenge dataset:
  - More than 40K multi-labeled shots
  - 101 concepts used as labels of the shots
  - Concepts cover a wide range of topics (e.g., indoor, outdoor, golf, baseball, T. Blair, etc.)
  - Dataset is partitioned in two independent datasets:
    - $\mathcal{D}$: 20K shots for training the concept detectors
    - $\mathcal{U}$: 492 shots referring to 5 sport events as the evaluation set (extracted from the rest 20K shots)

basketball	soccer	football	baseball	golf
119	198	71	53	51

BoW vocabulary & Trained concept detectors

- SIFT-based Bag-of-visual-words (BoW) vocabulary [1]
  - The vocabulary is learned using the dataset $\mathcal{D}$
  - From each shot keyframe we extract keypoints
  - These keypoints are described by 128-dimensional SIFT vectors
  - SIFT vectors are clustered to create a vocabulary of visual words

- To represent the i-th shot (in both $\mathcal{D}$ and $\mathcal{U}$) we extract keypoints and use the BoW vocabulary to derive the respective feature vector $\mathbf{s}_i$

- *K*=101 SVM-based concept detectors are trained using $\mathbf{s}_i$ in $\mathcal{D}$

$$\mathcal{G} = \{(d_\kappa(), h_\kappa),\ \kappa = 1, \ldots, K\}$$

[1] J. Molina, V. Mezaris, P. Villegas, et. al., "MESH participation to TRECVID2008 HLFE", Proc. TRECVID 2008 Workshop, November 2008, Gaithersburg, MD, USA.

Model vectors

- The trained concept detectors are applied in the dataset $\mathcal{U}$ to associate each shot feature vector $\mathbf{s}_i$ with a model vector

$$[x_{i,1}, \ldots, x_{i,K}]^T, \ \mathbf{x}_i \in \Re^K$$

  - Components of the model vector are in the range [0,1] expressing the DoC that a concept is depicted in the shot

$$x_{i,\kappa} = d_\kappa(\mathbf{s}_i)$$

Shot classification

- Concept-based (Max rule)
  - The test shot feature vector is derived using the BoW method
  - The trained concept detectors that correspond to the five event classes are evaluated $\acute{x}_{n,\kappa} = d_\kappa(\acute{\mathbf{s}}_n)$
  - The max rule is applied to classify the test shot

$$\acute{y}_n = \underset{\kappa \in [6,7,41,42,82]}{\arg\max} (\acute{x}_{n,\kappa})$$

- Event-based (NN classifier using model vectors)
  - The model vector of the test shot is derived
  - The NN classifier is applied to classify the test shot

$$\acute{y}_n = \underset{i \in [1,\ldots,N]}{\arg\max} (\mathbf{x}_i^T \acute{\mathbf{x}}_n)$$

Shot classification

- NN classifier in a discriminant subspace
  - The training model vectors are used to compute a linear projection matrix $\mathbf{W}_{LDA}$ or $\mathbf{W}_{SDA}$ using LDA or SDA respectively
  - The model vector of the test shot is derived and projected in the discriminant concept subspace

$$\acute{\mathbf{z}}_n = \mathbf{W}^T \acute{\mathbf{x}}_n$$

  - The NN classifier is applied in the discriminant subspace to classify the test shot

$$\acute{y}_n = \underset{i \in [1,\ldots,N]}{\arg\max} \left(\mathbf{z}_i^T \acute{\mathbf{z}}_n \right)$$

Evaluation

- Evaluation using 50-fold cross validation procedure:
  - At each fold
    - train set: 80%, test set: 20%
    - Correct classification rate (CCR) is retained
  - Average CCR (ACCR) is used as the performance measure

- Results

	Max Rule	Input Space	LDA	SDA
ACCR	60.5%	67.9%	63.2%	**69.4%**

Entity (face) classification

- Sheffield (UMIST) database:
  - Consists of 564 gray-scale cropped facial images of 20 subjects
  - Facial images cover a wide range of poses (multiview) from profile to frontal views as well as race, gender and appearance
- Preprocessing
  - Facial images are scaled to size 32x32 using bicubic interpolation
  - Scanned column-wise to form 1024-dimensional feature vectors

Face classification

- NN classifier in a discriminant subspace
  - Training feature vectors are used to compute linear projection matrices $\mathbf{W}_{PCA}$, $\mathbf{W}_{LDA}$, $\mathbf{W}_{SDA}$, $\mathbf{W}_{ISDA}$ using PCA, LDA, SDA or ISDA respectively
  - Feature vectors are projected in the discriminant subspace

$$\acute{\mathbf{z}}_n = \mathbf{W}^T \acute{\mathbf{x}}_n$$

  - The NN classifier is applied in the discriminant subspace to classify an unknown facial image

$$\acute{y}_n = \underset{i \in [1,\ldots,N]}{\arg\max} (\mathbf{z}_i^T \acute{\mathbf{z}}_n)$$

Evaluation

- Evaluation using 30-fold cross validation procedure:
  - At each fold
    - test set: 60%, train set: 40%
    - Correct classification rate (CCR) is retained
  - Average CCR (ACCR) is used as the performance measure

- Results

	PCA	LDA	SDA	ISDA
ACCR	94.9% (236)	95.5% (19)	97.2% (31)	**97.3**% (25)

Event-based video classification

- TRECVID 2010 Multimedia Event Detection (MED):
  - User-generated video clips (audio and video streams)
  - Annotation in terms of events (concept annotations are not provided)

- Events: “batting a run in”, “making cake”, “assembling shelter”
  - Development set: 1746 clips (~56 hours)
  - Evaluation set: 1742 clips (~59 hours)

	Target events			*Uninteresting events*			
	BR	*MC*	*AS*	*NBR*	*NMC*	*NAS*	*OT*
Train. set	50	48	48	4	12	3	1581
Eval. set	47	46	47	-	-	-	1602

BoW vocabulary & Trained concept detectors

- Independent dataset:
  - MediaMill dataset : 101 concepts, 40K shots
  - TRECVID 2010 SIN dataset: 130 concepts, 266K shots
- Construction of SIFT-based BoW vocabulary (extraction of keypoints, clustering of respective SIFT vectors)
- Using BoW each shot is represented with the respective feature vector
- *F*=231 SVM-based concept detectors are trained using the shot feature vectors

$$\mathcal{G} = \{(d_\kappa(), h_\kappa),\ \kappa = 1, \ldots, F\}$$

Video representation

- A video in 2010 TRECVID MED dataset is segmented to shots [1]
- Each shot is associated with a model vector using BoW vocabulary and trained concept detectors

$$\mathbf{x}_{p,q} = [x_{p,q,1}, \ldots, x_{p,q,K}]^T, \ \mathbf{x}_{p,q} \in \mathbb{R}^F$$

- A video is represented with a sequence of model vectors

$$\mathbf{X}_p = [\mathbf{x}_{p,1}, \ldots, \mathbf{x}_{p,l_p}], \mathbf{X}_p \in \mathbb{R}^{F \times l_p}$$

[1] E. Tsamoura, V. Mezaris, I. Kompatsiaris, "Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework", IEEE Int. Conf. on Image Processing, Workshop on Multimedia Information Retrieval, San Diego, CA, USA, October 2008, pp. 45-48.

Video classification

- Video similarity measure
  - Oriented median Hausdorff distance

$$D_H(\mathbf{Z}_t, \mathbf{Z}_p) = \underset{q}{\text{median}}(\min_{s} \parallel \mathbf{z}_{t,s} - \mathbf{z}_{p,q} \parallel)$$

  - Symmetric measure

$$d_H(\mathbf{Z}_t, \mathbf{Z}_p) = D_H(\mathbf{Z}_t, \mathbf{Z}_p) + D_H(\mathbf{Z}_p, \mathbf{Z}_t)$$

- Nearest neighbor rule

$$f(\mathbf{Z}_t) = \underset{p \in [1,\ldots,L]}{\text{argmin}} \left(d_H(\mathbf{Z}_t, \mathbf{Z}_p) \right)$$

Description of submitted runs

- IN
  - Classification of unknown videos using the NN-based Hausdorff distance directly in the 231-dimensional concept space
- LDA
  - LDA derives a discriminant subspace using the development set
  - Classification using NN-based Hausdorff distance in the LDA subspace
- SDA
  - SDA derives a discriminant subspace using the development set
  - Classification using NN-based Hausdorff distance in the SDA subspace
- ISDA1
  - ISDA derives a discriminant subspace using the development set
  - Classification using NN-based Hausdorff distance in the SDA subspace

Description of submitted runs

- ISDA2 – Training
  - 50-fold validation to produce event condidence scores
    - At each fold *development set* is split to 90% train set, 10% validation set
    - ISDA derives a discriminant subspace using train set
    - Validation videos are classified and a confidence score for each video and each event is retained
  - Construction of event probability models
    - The produced confidence scores are used to build a Gaussian distribution confidence score model for each event

Description of submitted runs

- ISDA2 – Testing
  - For each test video in the *evaluation set*, *C*=3 condidence scores are produced from its classification with the *C* events
  - The confidence scores are weighted with the probability that these scores belong to the respective event (using the event probability models)
  - The maximum weighted score is taken to indicate the underlying event

Description of submitted runs

- ISDA3
  - Same procedure as ISDA1 (derive and project test video in the ISDA subspace),
  - Use windowing version of Hausdorff distance to classify test video
    - For comparing test video $\mathbf{Z_t}$ with train video $\mathbf{Z_p}$ a sliding window of length $\min\{l_t, l_p\}$ is used to produce $m_p = |l_t + l_p| + 1$ Hausdorff distance values

$$d_p^{i_p}, \;\; i_p = 1, \ldots, m_p$$

    - This is done with all train videos in order to classify the test video

$$\operatorname{argmin}(d_p^{i_p}), \;\; p = 1, \ldots, L, \;\; i_p = 1, \ldots, m_p$$

- ISDA4
  - As in ISDA3 a windowing version of Hausdorff distance is used
  - As in ISDA2 event probability models are constructed and confidence scores are weighted with the probability that these scores belong to the respective event

Normalization detection cost (NDC)

- Linear combination of missed detection (MD) and false alarm (FA) probabilities

$$\mathrm{NDC}^i = C^i_{MD} P^i_{MD} P^i_T + C^i_{FA} P^i_{FA} (1 - P^i_T)$$

$$P^i_{MD} = \frac{N^i_{MD}}{N^i_T} \qquad P^i_{FA} = \frac{N^i_{FA}}{N^i_T}$$

  - Provided: $P^i_T, C^i_{MD}, C^i_{FA}$

- A small NDC shows a good performance

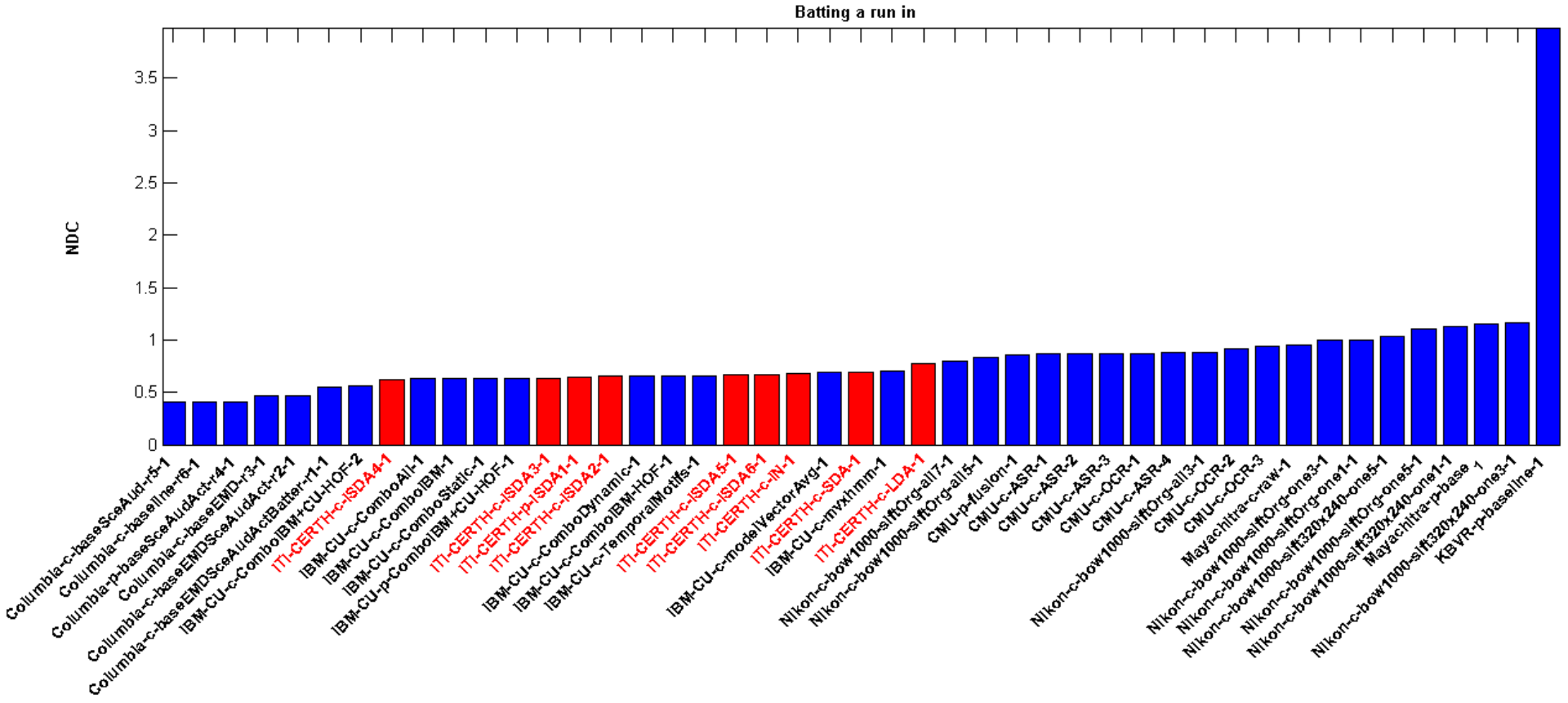
TRECVID 2010 MED Results

- Good results regarding BR event (several relevant concepts detectors in our pool, e.g., “grass”, “running”, etc.), and moderate good results for the other two events (although few relevant concept detectors in our pool)
- SDA and ISDA methods in general perform better than LDA and IN
- ISDA performs better than SDA for two out of three events
- For BR event, windowing Hausdorff distance as well as its combination with event probability models improved performance

Method	Batting a run in				Assembling a shelter				Making a cake			
	TP	*FA*	*NDC*	*D*	*TP*	*FA*	*NDC*	*D*	*TP*	*FA*	*NDC*	*D*
IN	32	48	0.6766	231	13	52	1.1044	231	12	36	1.0127	231
LDA	21	30	0.7766	6	6	24	1.0482	6	6	43	1.1925	6
SDA	27	36	0.6936	90	11	28	**0.9692**	90	5	15	0.9979	90
ISDA1	29	35	0.6436	42	9	38	1.0871	42	6	15	**0.9840**	42
ISDA2	29	37	0.6584	42	10	41	1.0877	42	11	33	1.0117	42
ISDA3	28	31	0.6351	42	11	45	1.0958	42	5	21	1.05	42
ISDA4	29	32	**0.6213**	42	11	49	1.1255	42	8	59	1.2691	42

TRECVID 2010 MED Results

- For event BR, best performance is achieved among all submissions that use only visual information (opposed to submissions exploiting audiovisual information)



Conclusions

- ✓ We propose organizing information around events for closing the semantic gap between human and machine interpretations
- Joint content-event model
  - Provides a content model to represent multimedia content
  - Treats events as first class entities (uses multimedia data to provide the experiential dimension of pre-existing events)
  - Uses a referencing mechanism to promote automatic enrichment of event elements with information extracted from multimedia content
- Referencing mechanism
  - Model vector approach significantly reduces training time as no additional time is introduced for learning BoW vocabulary and training event/concept detectors specifically for the dataset of interest
  - Processing in a reduced concept subspace, further reduces computation time, offers lower storage requirements, and improved classification accuracy
  - Promising results on Mediamill challenge dataset (shot classification), Sheffield dataset (face recognition), TRECVID 2010 MED competition (video classification)

Publications

- N. Gkalelis, V. Mezaris, I. Kompatsiaris, "Mixture subclass discriminant analysis", IEEE Signal Processing Letters, accepted for publication, 2011.
- N. Gkalelis, V. Mezaris, I. Kompatsiaris, "High-level event detection in video exploiting discriminant concepts", Proc. 9th International Workshop on Content-Based Multimedia Indexing (CBMI 2011), Madrid, Spain, June 2011.
- I. Tsampoulatidis, N. Gkalelis, A. Dimou, V. Mezaris, I. Kompatsiaris, “High-level Event Detection System Based on Discriminant Visual Concepts”, Proc. ACM International Conference on Multimedia Retrieval (ICMR 2011), Trento, Italy, April 2011.
- A. Moumtzidou, A. Dimou, N. Gkalelis, S. Vrochidis, V. Mezaris, I. Kompatsiaris, "ITI-CERTH participation to TRECVID 2010", Proc. TRECVID 2010 Workshop, November 2010, Gaithersburg, MD, USA.
- N. Gkalelis, V. Mezaris, I. Kompatsiaris, "Automatic event-based indexing of multimedia content using a joint content-event model", Proc. ACM Multimedia 2010, Events in MultiMedia Workshop (EiMM10), Firenze, Italy, October 2010.
- N. Gkalelis, V. Mezaris, I. Kompatsiaris, "A joint content-event model for event-centric multimedia indexing", Proc. Fourth IEEE International Conference on Semantic Computing (ICSC 2010), Pittsburgh, PA, USA, September 2010, pp. 79-84.

Thank you for your attention!

Questions?