# The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks

Stefanie Nowak

ImageCLEF 2011 Lab Presentation

Amsterdam, 20.09.2011

Outline

- Task Description

- Ground Truth Assessment

- Performance Measures

- Participation

- Results

- Conclusions

Fraunhofer
IDMT

Task Description

- 2 subtasks
    - 1) Image Annotation Task
    - 2) Concept-based Image Retrieval Task (NEW!)
- Test collection:
    - Flickr photos based on interestingness
      (MIR Flickr Set + MIR Flickr 1 Million Set)
- 3 Configurations:
    1) Textual information (EXIF tags, Flickr User Tags)
    2) Visual information (photos)
    3) Multi-modal information (all)

Fraunhofer

IDMT

Task Description: Annotation Task

- Automated annotation of 99 visual concepts in photos

- Multi-label Scenario

- Test collection:

  - Trainingset: 8,000 photos + Flickr User Tags + EXIF data + GT

  - Testset: 10,000 photos + Flickr User Tags + EXIF data

- Evaluation Methodology:

  - Interpolated Average Precision (iAP)

  - Example-based F-measure (F-ex)

  - Semantic R-Precision (SR-Prec) ($\rightarrow$ see later)

- Fully assessed test collection

Fraunhofer
IDMT

Task Description: Retrieval Task

- Retrieval of images based on certain topics

- 40 topics + example photos

- Test collection:

  - Trainingset of annotation task (including GT for 99 concepts)

  - Testset: 200,000 photos + Flickr User Tags + EXIF data

- Performance Measures:

  - non-interpolated Average Precision (AP),

  - Precision@X (P@10, P@20, P@100)

  - Concept-based R-Precision (R-Prec)

- Pooling depth: 100 images per run for each topic

Fraunhofer
IDMT

Objectives & Challenges

- Exploitation of different knowledge sources
- Benefit of annotation approaches in retrieval scenarios
- Abilities to predict subjective concepts (e.g. sentiments)

- Challenges:
  - Varying number of labels per image
  - Unbalanced amount of data per concept
  - Diversity of image content per concept
  - Missing user tags and EXIF data
  - Different qualities of image metadata

Fraunhofer
IDMT

Visual Concepts

- **Reuse of concepts and labels**
  - 49 (out of 52) concepts from 2009
  - 41 concepts from 2010
- **9 novel sentiment concepts:**

Category	Subcategory	Number concepts	Percentage
Content Element	Landscape Elements	12	44.44%
	Urban Elements	3	
	General	3	
	Persons & person related	12	
	Vehicle	7	
	Animals	7	
Scene Description	Abstract	9	23.23%
	Activity	1	
	Seasons	4	
	Place	2	
	Daytime	4	
	Events	3	
Representation	General	3	14.14%
	Illumination	4	
	Art	3	
	abstract	4	
Quality	Blurring	4	6.06%
	Aesthetics	2	
Emotion and Affect	opinions	4	12.12%
	emotions	8	

Retrieval Task: Topics

Number	Title	Source
1	Graffiti on buildings or walls	WR 11
2	Toy vehicle	
3	Single person doing sports on the sea	WR 09/10
4	Airplane in the sky	WR 10
5	Rider on horse	WR 09/10

<topic>
<number>3</number>
<title>single person doing sports on the sea</title>
<image>{6030CEAE-F5BF-4FB1-A188-331793DB9C13}.jpg</image>
<image>…</image>
<narrative>We like to find photos of one person doing sports on the sea.
Pictures with several persons doing sports are not relevant, neither are
only surf boards or other sport equipment.
Also persons doing sports at the beach are not relevant.</narrative>
</topic>

Fraunhofer
IDMT

GT Assessment: Annotation Task

- 1 HIT = 9 images +1 gold image
- Automated completeness check
- (Internal) rejection criteria:
  - Failed gold standard test
 - deviation of >90°
- GT: Majority vote of 5 opinions

What sentiments does this image convey?

Please choose at least one sentiment.

- active
- euphoric
- scary
- funny
- unpleasant
- happy
- melancholic
- calm/comforting ☑
- inactive

ACTIVE, MISERY, PLEASURE, nothing, PASSIVE

What triggered that sentiment the most?

mostly overall impression

motif — overall impression

Example

Batch	# HITs	Distinct turkers	Completion time	Reward	# votes
Pretest	120	22	3 min, 12 sec	0.05$	3
Training	4225	258	2 min, 36 sec	0.07$	5
Test I	2745	156	2 min, 8 sec	0.07$	5
Test II	2815	(both parts)	1 min, 44 sec	0.07$	5

Fraunhofer
IDMT

GT Assessment: Retrieval Task

- 1 HIT = 22 images + 2 gold standard images (relevant + irrelevant)
- Number of HITs depending on topic
- Reward per HIT: 0.03$

- 3 votes per image
- Min 5, max 41 distinct turkers per topic
- Completion time: 31 sec - 1min19sec

- Rejection criteria: gold standard test
- Postscreening procedure

Topic: *Fish in water*

GT Assessment: Retrieval Task

What is the HIT about?

In the following HIT you will be given 24 images that were assigned to a certain topic. Please select any image that you think does not fit the topic and tell us the reason why you want to eliminate it.

Guidelines:

- Choose any image that does not fit the given topic.
- For any selected image, use the textbox to tell us why you chose it.

NOTE: You can only submit your answers if all of these steps are fulfilled. Incomplete answers are highlighted by a red border during submission.

Please be advised that occasionally there might be a small number of adult or disturbing images despite our effort to filter them out.

How are you paid?

You will be paid the amount that is defined in the HIT. Your submission will typically be approved/rejected within 7 days.

Topic: Bridges Not Over Water

We would like to find photos of all kinds of bridges, but not over water. Bridges over a valley, between buildings and so on are relevant. Half finished and half-destroyed bridges are relevant as long as it is evident that they will be or that they were bridges.

Examples:

Fraunhofer
IDMT
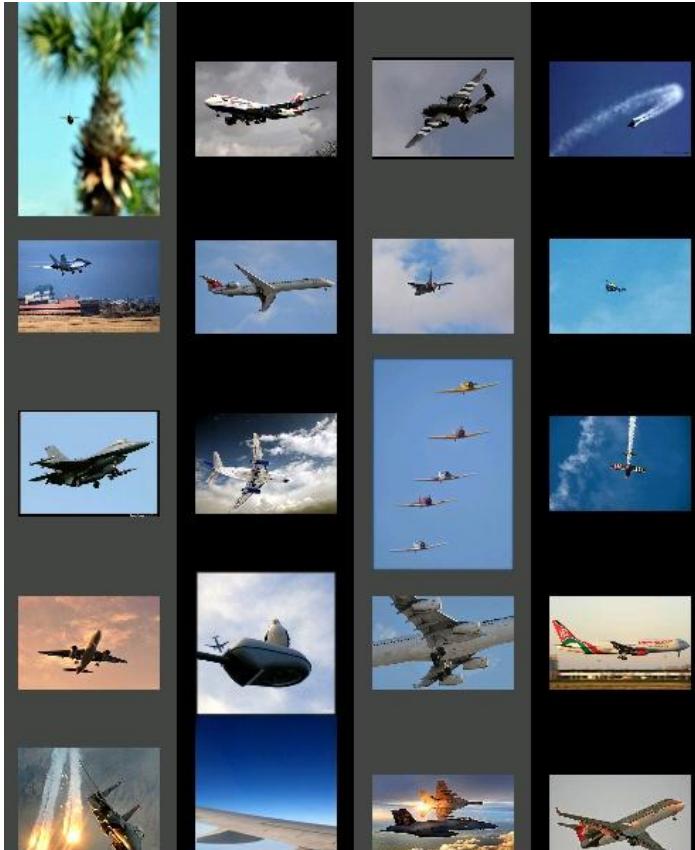
GT Assessment: Retrieval Task

| Topic	HITs	Images	Pool	turkers	time	Topic	HITs	images	Pool	turkers	time
1	189	1365	Master	8	1 min 9 s	21	195	1412	Master	9	39 s
2	200	1398	Usual	17	1 min 2 s	22	210	1531	Master	13	42 s
3	153	1109	Usual	11	46s	23	159	1155	Master	13	48 s
4	199	1315	Usual	22	43 s	24	192	1396	Master	13	40 s
5	201	1442	Usual	14	41 s	25	201	1465	Master	9	31 s
6	227	1477	Usual	22	52 s	26	207	1516	Master	10	39 s
7	195	1416	Usual	35	1 min 8 s	27	156	1142	Master	11	35 s
8	186	1346	Usual	25	58 s	28	183	1339	Master	6	1 min 4 s
9	183	1331	Usual	32	1 min 7 s	29	198	1450	Master	5	47 s
10	192	1389	Usual	20	49 s	30	189	1365	Master	7	44 s
11	207	1508	Usual	26	51 s	31	192	1402	Master	13	57 s
12	162	1184	Usual	41	1 min 16 s	32	222	1616	Master	8	41 s
13	231	1688	Master	10	46 s	33	183	1335	Master	10	38 s
14	162	1186	Master	13	59 s	34	219	1599	Master	7	40 s
15	147	1071	Master	9	1 min 19 s	35	222	1610	Master	9	40 s
16	204	1489	Master	14	44 s	36	216	1577	Master	9	30 s
17	177	1294	Master	8	43 s	37	228	1659	Master	15	45 s
18	267	1943	Master	11	34 s	38	234	1716	Master	9	33 s
19	159	1152	Master	8	56 s	39	210	1531	Master	6	44 s
20	210	1536	Master	9	45 s	40	201	1454	Master	12	42 s
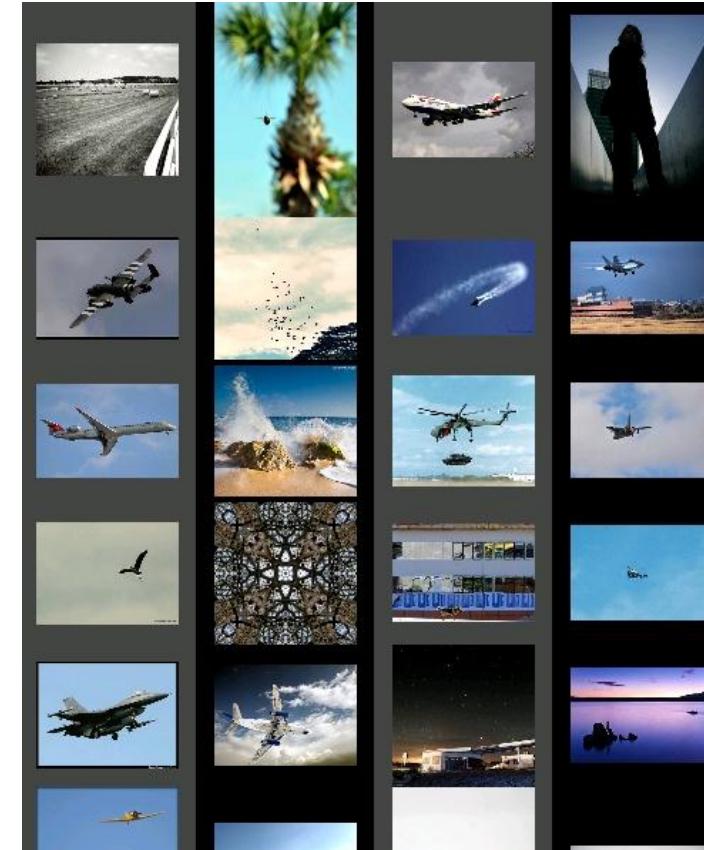
- Rejection: 8.7% of all HITs
- Master rejection: 7.5% of all HITs posed to master workers
- Usual worker rejection: 10.5% of all HITs posed to usual workers

GT Assessment: Retrieval Task

Topic 4: airplane in the sky – normal pool
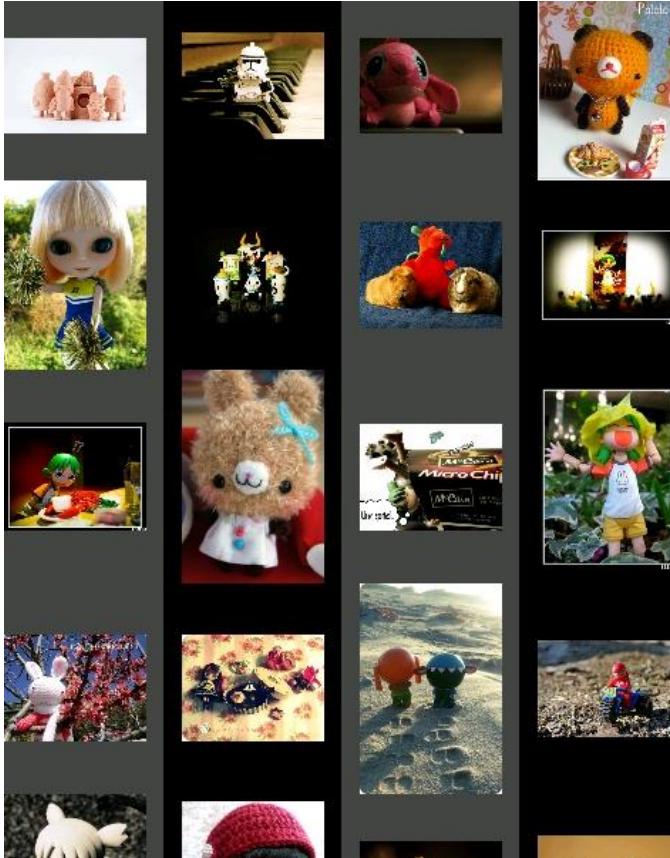

at least 2 votes
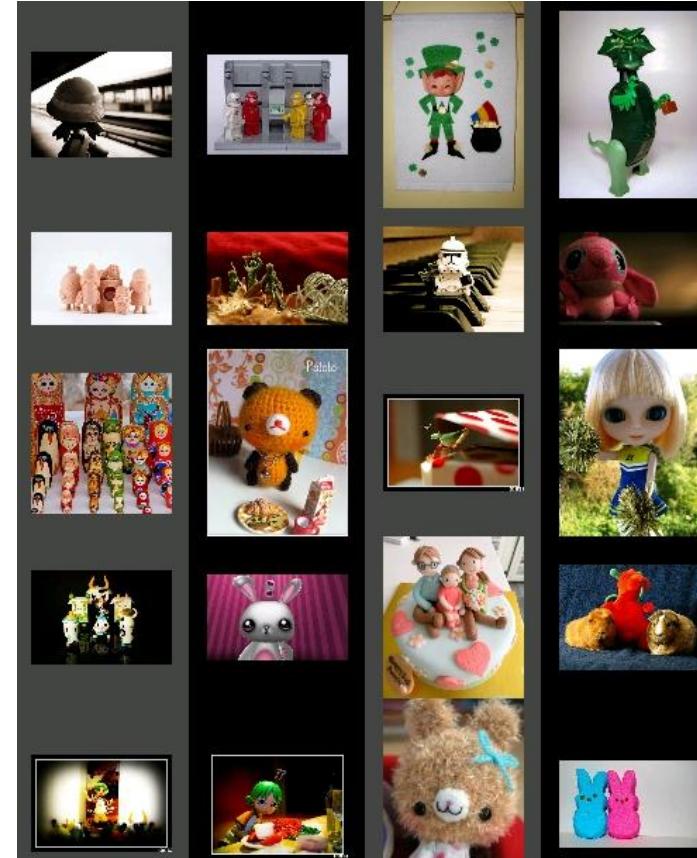
at least 1 vote

GT Assessment: Retrieval Task

Topic 30: cute toys arranged to a still-life – master pool

at least 2 votes
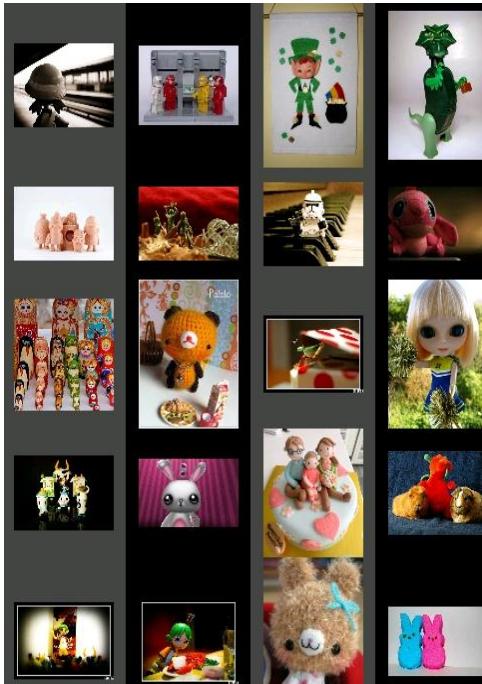
at least 1 vote

GT Assessment: Retrieval Task

at least 2 votes

at least 1 vote
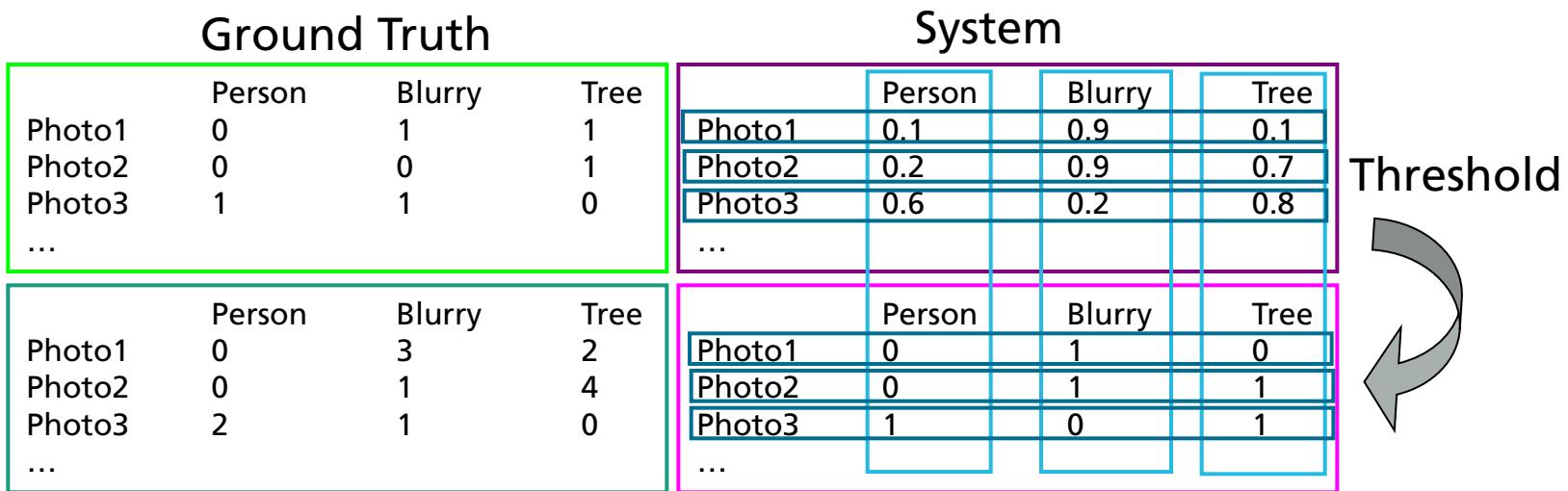
Votes:	1	2	3
	495	360	214
	296	86	36
	414	99	22
	179	78	51
	476	100	54
	277	106	72
	543	139	36
	140	76	43
	354	167	81
	194	136	96
	478	182	100
	657	274	90
	101	52	19
	238	127	61
	358	206	118
	331	211	113
	366	194	84
	24	15	6
	531	213	78
	193	118	54
	210	29	13
	657	152	27
	147	68	33
	136	57	24
	208	116	63
	146	37	9
	242	149	76
	146	132	105
	125	94	61
	308	217	114
	234	110	44
	611	309	92
	164	92	42
	415	338	215
	479	457	415
	227	99	29
	340	214	126
	152	69	24
	214	121	38
	483	312	151

Visual impression

Fraunhofer
IDMT

Performance Measures
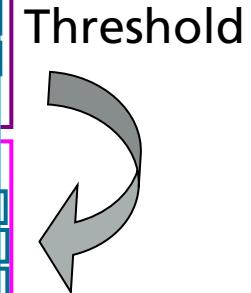
- Evaluation per concept vs. evaluation per media item
- Ranked predictions vs. binary decisions
- Binary relevance vs. graded relevance

Ground Truth

	Person	Blurry	Tree
Photo1	0	1	1
Photo2	0	0	1
Photo3	1	1	0
...			

	Person	Blurry	Tree
Photo1	0	3	2
Photo2	0	1	4
Photo3	2	1	0
...			

System

	Person	Blurry	Tree
Photo1	0.1	0.9	0.1
Photo2	0.2	0.9	0.7
Photo3	0.6	0.2	0.8
...			

Threshold

	Person	Blurry	Tree
Photo1	0	1	0
Photo2	0	1	1
Photo3	1	0	1
...			
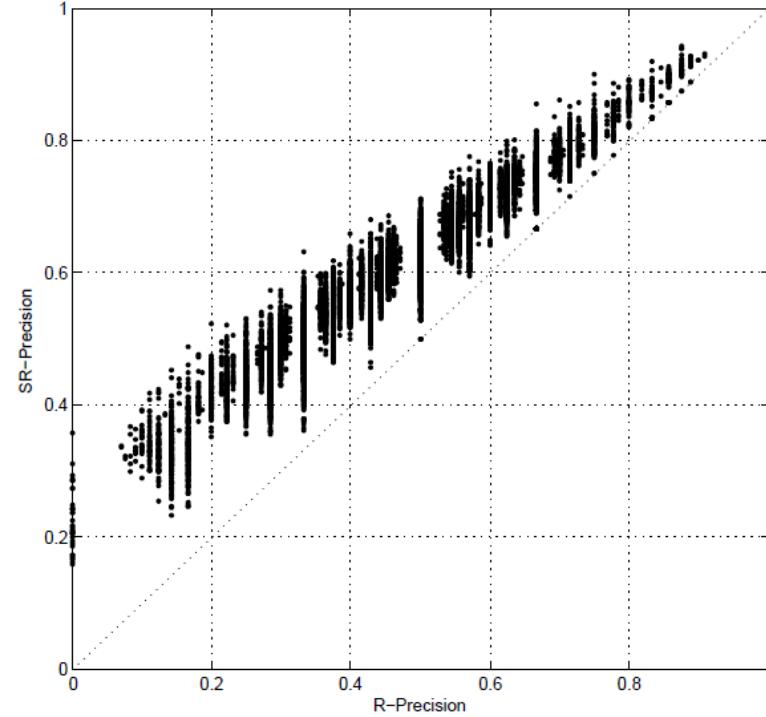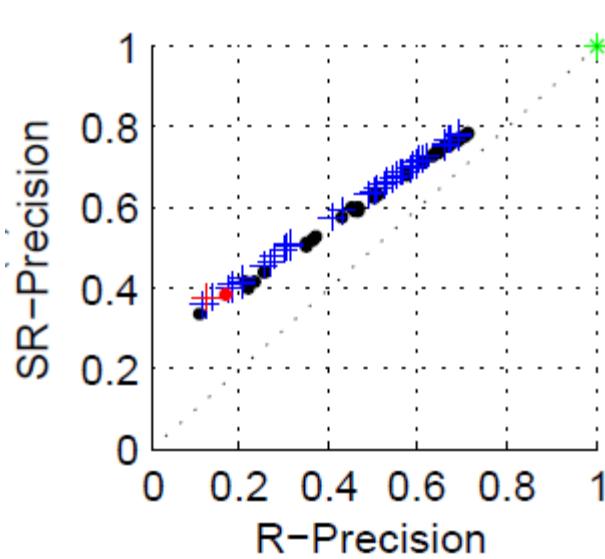
Performance Measures

	Concept-based	Example-based
Binary predictions		F-Ex
Ranked predictions	Average Precision (P@10, P@20, P@100, R-Precision)	SR-Precision

Ground Truth

	Person	Blurry	Tree
Photo1	0	1	1
Photo2	0	0	1
Photo3	1	1	0
…			

	Person	Blurry	Tree
Photo1	0	3	2
Photo2	0	1	4
Photo3	2	1	0
…			

System

	Person	Blurry	Tree
Photo1	0.1	0.9	0.1
Photo2	0.2	0.9	0.7
Photo3	0.6	0.2	0.8
…			

Threshold

	Person	Blurry	Tree
Photo1	0	1	0
Photo2	0	1	1
Photo3	1	0	1
…			

Fraunhofer
IDMT

Performance Measure: Semantic R-Precision

- Extension of example-based R-Precision

- Flickr Tag Similarity to determine semantic relatedness

Participation: Annotation Task

- 48 groups registered
- 42 teams signed licence agreement

- 18 groups from 11 countries participated with 79 runs
- Max 5 runs per team

Team	Textual	Visual	Multi-modal
BPACAD	1	1	3
BUFFALO	-	5	-
CAEN	-	4	-
CEALIST	1	1	3
DBIS	-	5	-
HHI	-	5	-
IDMT	1		4
ISIS	-	3	2
LAPI	-	2	-
LIRIS	2	1	2
MEIJI	1	2	2
MLKD	1	1	3
MRIM	-	3	1
MUFIN	-	-	4
NII	-	5	-
REGIMVID	1	-	-
TUBFI	-	4	1
UNIKLU	-	4	-
Total:	**8**	**45**	**26**

Fraunhofer
IDMT

Participation: Retrieval Task

Team	Textual	Visual	Multi-modal	Automated	Manual
ISIS	-	10	-	3	7
MEIJI	2	2	6	10	-
MLKD	4	2	4	2	8
REGIMVID	1	-	-	1	-
Total	7	14	10	16	15

- Max 10 runs per team

- 4 teams submitted a total of 31 runs

- Runs consider all configurations

Fraunhofer

IDMT

Results: Annotation Task

Team	MiAP	Conf
TUBFI	0.443	M
LIRIS	0.437	M
BPACAD	0.436	M
ISIS	0.433	M
MLKD	0.402	M
CEALIST	0.384	M
CAEN	0.382	V
MRIM	0.377	M
IDMT	0.371	M
NII	0.337	V
HHI	0.335	V
MEIJI	0.304	T
MUFIN	0.299	M
BUFFALO	0.249	V
DBIS	0.230	V
UNIKLU	0.207	V
REGIMVID	0.204	T
LAPI	0.177	V

Team	F-Ex	Conf
ISIS	0.622	M
CAEN	0.600	V
BPACAD	0.593	M
HHI	0.588	V
LIRIS	0.576	M
TUBFI	0.566	M
MLKD	0.560	V
MRIM	0.552	M
IDMT	0.552	M
BUFFALO	0.527	V
DBIS	0.518	V
CEALIST	0.508	M
MEIJI	0.495	M
UNIKLU	0.469	V
MUIN	0.462	M
LAPI	0.390	V
NII	0.298	V
REGIMVID	0.141	T

Results: Annotation Task - Configurations

■ Visual Configuration

Team	MiAP
TUBFI	0.388
CAEN	0.382
ISIS	0.375
BPACAD	0.367
LIRIS	0.355
NII	0.337
MRIM	0.336
HHI	0.335
MLKD	0.311
CEALIST	0.301
BUFFALO	0.249
DBIS	0.230
UNIKLU	0.207
MEIJI	0.204
LAPI	0.177

Team	F-Ex
ISIS	0.612
CAEN	0.600
HHI	0.588
BPACAD	0.568
MLKD	0.560
TUBFI	0.552
MRIM	0.544
LIRIS	0.539
BUFFALO	0.527
DBIS	0.518
CEALIST	0.503
MEIJI	0.472
UNIKLU	0.469
LAPI	0.390
NII	0.298

■ Textual Configuration

Team	MiAP
BPACAD	0.346
IDMT	0.326
MLKD	0.326
LIRIS	0.321
MEIJI	0.304
CEALIST	0.292
REGIMVID	0.204

Team	F-Ex
IDMT	0.525
MLKD	0.506
BPACAD	0.502
CEALIST	0.479
MEIJI	0.459
LIRIS	0.432
REGIMVID	0.141

Fraunhofer
IDMT

Results: Annotation Task - Configurations

◼ Visual Configuration

Team	MiAP
TUBFI	0.388
CAEN	0.382
ISIS	0.375
BPACAD	0.367
LIRIS	0.355
NII	0.337
MRIM	0.336
HHI	0.335
MLKD	0.311
CEALIST	0.301
BUFFALO	0.249
DBIS	0.230
UNIKLU	0.207
MEIJI	0.204
LAPI	0.177

Team	F-Ex
ISIS	0.612
CAEN	0.600
HHI	0.588
BPACAD	0.568
MLKD	0.560
TUBFI	0.552
MRIM	0.544
LIRIS	0.539
BUFFALO	0.527
DBIS	0.518
CEALIST	0.503
MEIJI	0.472
UNIKLU	0.469
LAPI	0.390
NII	0.298

◼ Textual Configuration

Team	MiAP
BPACAD	0.346
IDMT	0.326
MLKD	0.326
LIRIS	0.321
MEIJI	0.304
CEALIST	0.292
REGIMVID	0.204

Team	F-Ex
IDMT	0.525
MLKD	0.506
BPACAD	0.502
CEALIST	0.479
MEIJI	0.459
LIRIS	0.432
REGIMVID	0.141

◼ MiAP:
close results in both configurations

Fraunhofer
IDMT

Results: Annotation Task - Configurations

■ Visual Configuration

Team	MiAP
TUBFI	0.388
CAEN	0.382
ISIS	0.375
BPACAD	0.367
LIRIS	0.355
NII	0.337
MRIM	0.336
HHI	0.335
MLKD	0.311
CEALIST	0.301
BUFFALO	0.249
DBIS	0.230
UNIKLU	0.207
MEIJI	0.204
LAPI	0.177

Team	F-Ex
ISIS	0.612
CAEN	0.600
HHI	0.588
BPACAD	0.568
MLKD	0.560
TUBFI	0.552
MRIM	0.544
LIRIS	0.539
BUFFALO	0.527
DBIS	0.518
CEALIST	0.503
MEIJI	0.472
UNIKLU	0.469
LAPI	0.390
NII	0.298

■ Textual Configuration

Team	MiAP
BPACAD	0.346
IDMT	0.326
MLKD	0.326
LIRIS	0.321
MEIJI	0.304
CEALIST	0.292
REGIMVID	0.204

Team	F-Ex
IDMT	0.525
MLKD	0.506
BPACAD	0.502
CEALIST	0.479
MEIJI	0.459
LIRIS	0.432
REGIMVID	0.141

■ MiAP:
close results in both configurations

■ F-Ex:
significant differences

Fraunhofer
IDMT

Results: Annotation Task

	Textual	Visual	Multi-modal
MiAP	0.346	0.388	0.443
F-Ex	0.525	0.612	0.622
SR-Precision	0.677	0.734	0.742
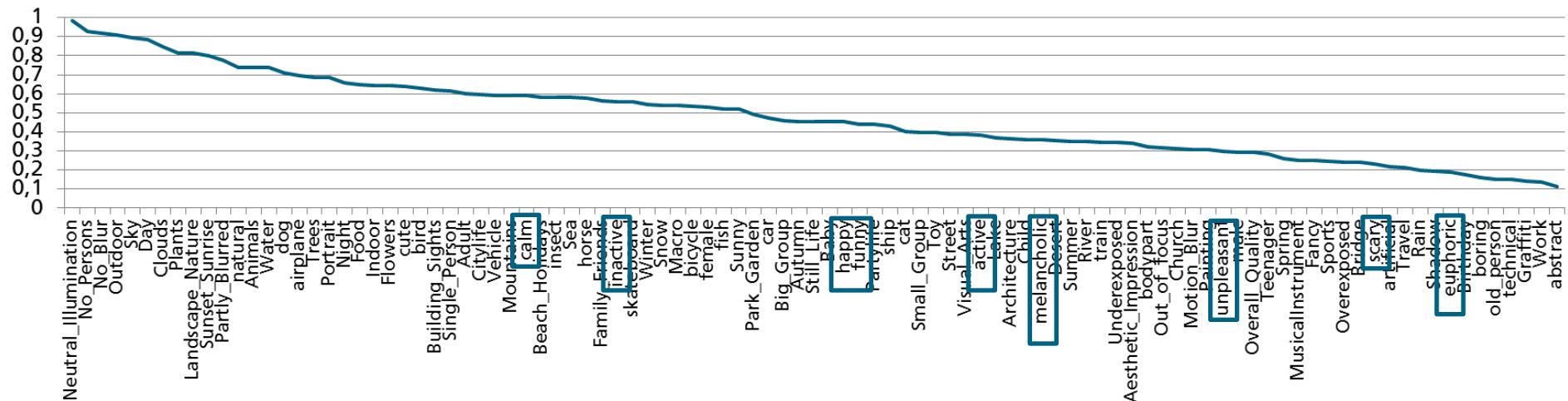
- Multi-modal runs:
    - outperform visual and textual ones
    - close results to visual ones in example-based evaluation:
- Textual and visual runs:
    - Close in concept-based evaluation
   - Significant differences in example-based evaluation

Fraunhofer

IDMT

Results per concept - Annotationtask

- Ø 0.48 MiAP per concept (all configurations and teams)
- 99 concepts:
    - 79 best by multi-modal configuration
    - 17 best by visual approach
    - 3 concepts best by textual approach

- Best concepts:
  Neutral-Illumination, No-Persons, No-Blur, Outdoor, Sky, Day, Clouds, Plants
- Worst concepts:
  abstract, work, Graffiti, technical , old-person , boring
- → almost identical to last year

Fraunhofer
IDMT

Results per concept - Annotationtask

■ Best results by any team per concept sorted by AP

■ Detection performance of sentiments
Calm → inactive → happy → funny → active → melancholic → unpleasant → scary → euphoric

Fraunhofer
IDMT
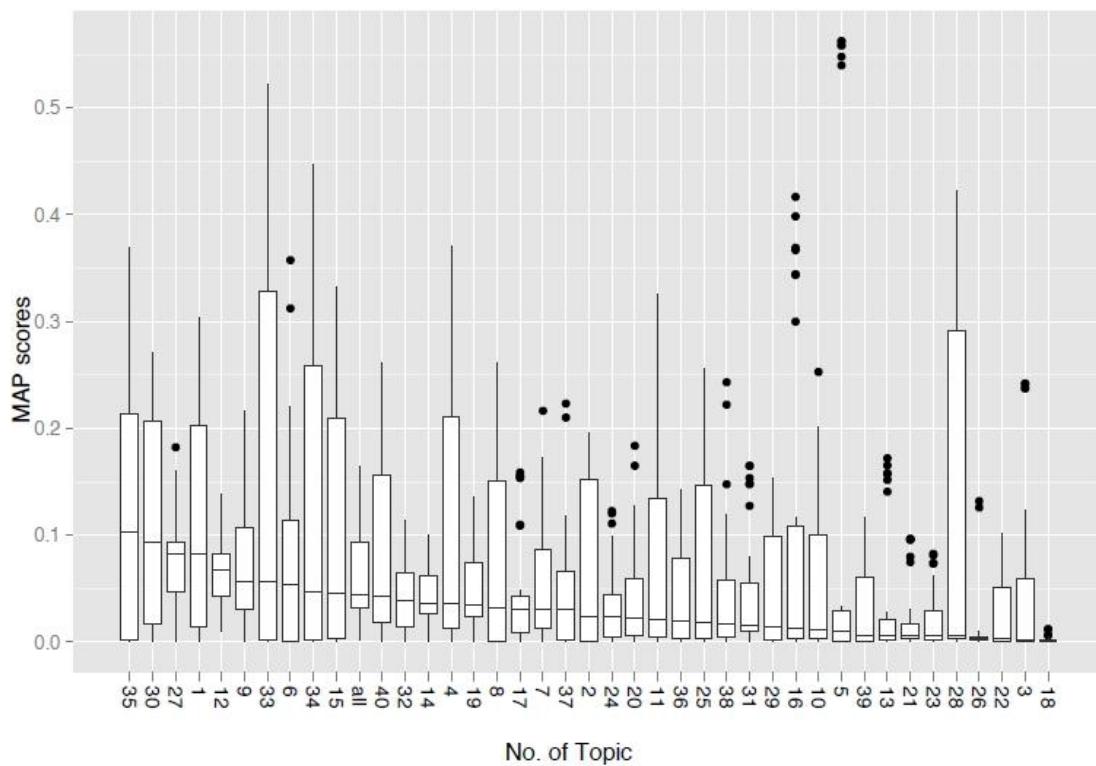
Results: Retrieval Task

Run	MAP	P@10	P@20	P@100	R-Prec	Automation
Visual						
ISIS	0.0997	0.3125	0.3050	0.2428	0.1712	Manual
ISIS	0.0430	0.1675	0.1550	0.1270	0.0974	Automated
MLKD	0.0361	0.1525	0.1375	0.1080	0.0883	Manual
MEIJI	0.0017	0.0150	0.0150	0.0197	0.0151	Automated
Textual						
MLKD	0.1546	0.4100	0.3838	0.3102	0.2366	Manual
MLKD	0.0849	0.3000	0.2800	0.2188	0.1530	Automated
MEIJI	0.0227	0.0900	0.0962	0.0865	0.0628	Automated
REGIMVID	0.0042	0.0650	0.0550	0.0352	0.0200	Automated
Multimodal						
MLKD	0.1640	0.3900	0.3700	0.3180	0.2467	Manual
MEIJI	0.0444	0.1625	0.1650	0.1465	0.1053	Automated

- Performance for automated runs half as good as for manual ones

Fraunhofer
IDMT

Results: Retrieval Task

MAPs, sorted by descending Median of single Topics

No. of Topic

Well performing topics:
- *close-up of bird(s)*
- *cute toys arranged to a still-life*
- *family holidays at the beach in summer*

Topics not working:
- *female old person*
- *single person doing sports on the sea*
- *portrait that is out of focus*

Fraunhofer
IDMT

Conclusions

- Strong participation in annotation task
    - 18 teams from 11 countries
    - Textual approaches significantly increase
    - Introduction of sentiment concepts
- New concept-based retrieval task
    - Challenging task
    - Rather low participation
- Ground Truth obtained with Amazon Mechanical Turk

- Results for all runs: http://imageclef.org/2011/photo

Fraunhofer IDMT

Photo Annotation Task 2012

- Are you interested in the concept-based retrieval task?
    - Problems?
    - Reasons not to particpate?
    - Too difficult?

- Annotation Task test collection was used for 3 years now
    - Interest in different collection?
    - What collection characteristics are important?

- Other/further suggestions?

Thank you very much.

Questions?

Stefanie Nowak
Audio-Visual Systems
Fraunhofer IDMT
www.idmt.fraunhofer.de

Fraunhofer
IDMT