
TEACHING STUDENTS HOW TO TEACH MACHINES

A problem-based learning activity on image classification

- Mohedano-Robles, Eva

Technical University of Catalonia (UPC)

Signal Theory and Communications Department

Colom 1, 08222 Terrassa (Catalonia)

evamohe@gmail.com

- Giro-i-Nieto, Xavier

Technical University of Catalonia (UPC)

Signal Theory and Communications Department

Colom 1, 08222 Terrassa (Catalonia)

xavier.giro@upc.edu

1. **ABSTRACT:**

This paper presents a Problem-Based Learning activity that introduce machine learning contents in a study plan that is based on electronics and signal processing subjects. The activity proposes students to follow the same evaluation protocols adopted in scientific challenges, where differents research groups test their own machine learning techniques on a common dataset and evaluation metrics. Teams of students adopt the role of a research group to develop their novel solution for classifying images depending on

whether they represent a certain semantic concepts. Student must implement software solutions that are tested and presented in class to prove their performance and originalty. This paper includes a detailed evaluation of the workload and satisfaction activity, based on questionaries answered by the students.

KEYWORDS in English**:** MACHINE LEARNING, PROBLEM-BASED LEARNING, GROUPS

2. FIELD OF KNOWLEDGE:

- Engineering and Architecture

3. SUBJECT AREA:

- Autonomous Student Learning

The Scientific Committee reserves the right to decide the final area of the proposals.

4. PRESENTATION CATEGORY:

- Oral Presentation

The Scientific Committee reserves the right to decide the final presentation category.

5. DEVELOPMENT: 25.000 – 35.000 characters (with spaces)

a) Objectives

Machine learning, a core part of the artificial intelligence technologies, has gained a growing interest from the scientific community for the past decades [Bishop06]. The possibility of designing machines that auto-configure just through the observation of acquired data has boosted the research of solid mathematical foundations which are programmed and rapidly computed by microprocessors. The evolution of the digital cameras has driven into a situation where large amount of data are generated and that require an automated analysis to filter and interpret their contents. These data, whose nature is completely perceptual, have been combined with the cited learning algorithms thanks to the advances in the field of signal processing. Nowadays, several automated systems are capable to provide an interpretation of audio, images and videos thanks to the maturity of the three fields: data acquisition, signal processing and machine learning. For example, these technologies have been applied to develop techniques for optical character recognition, face and smiles detectors in digital cameras, or applications that identify songs for mobile devices.

During the reformulation of the study plans of the Electrical Engineering degree in the Engineering School of Terrassa (UPC, Catalonia), it was detected that the previous study plan gave special attention to the electronic data acquisition and signal processing of audio and video signals, but that it lacked any formal explanation about the posterior explotation through machine learning techniques. This situation was reasonable as, traditionally, these have been the two basic pilars in the Electrical Engineering studies.

However, given the advances and redefinition of the study plan into a more generic "Degree in Audiovisual Systems", it was decided to introduce some basic machine learning contents in the programme.

This paper reports on a novel learning activity that aims at introducing machine learning contents in an Audio and Video Technologies study plan. The proposed methodology follows a *Problem Based Learning (PBL)* [Doud99] approach inspired in the actual problems solved by the researchers in the field. It was observed that this scientific community has defined very clear and evaluable strategies to test their approaches. There exist several evaluation campaigns whose completion require an understanding of the main concepts involved in machine learning [Smeaton09]. Moreover, their practical nature and transparent evaluation makes them very convenient to be used in a course to drive the learning experience of the students. The main contribution of this activity is a mapping of the classic stages in the definition and solving of a machine learning problem into a set of active learning activities aimed at solving a practical problem.

The presented activity also targets at three more generic skills which are included as learning goals in the study plan. Firstly, the solving of the problem requires an important degree of autonomous learning as the student will need to use external sources to solve the problem. Secondly, given the volume of the task and available workload, the activity is planned to be solved in a group. In fact, the whole activity is inherently presented as a competition, although the students are advised from the beginning that their final mark is based on the acomplishments of certain tasks, not on the relative performance of their works. Finally, the solutions developed by every team must be orally defended in front of their class mates as well as documented and publicly shared. This requirement includes communication skills, both oral and written, in the list of competences developed with the activity.

b) Description of the activity

The goal of the students is to design, develop and test a software capable of giving a semantic interpretation to an image. The activity is designed in six different stages that follow the traditional development in the design of a machine learning solution. There are basically two parts, a first basic one where the students focus in developing their programming skills and understanding the evaluation process, and a second one where they must introduce the *learning* capability into their solution. Each of the two parts include an oral exposition in class of the work completed. The six stages are the following:

<STAGE 1> Introduction and evaluation metrics: A regular session in class driven by the teacher describes the basic problem as the need to detect if a certain semantic concept is present in a digital image. The discussion is taken so that the students realise that in order to test the quality of this system they will need to design an experiment with several images. Each image will be automatically processed and the decision taken by the software will be compared with a ground truth. The teacher will introduce the two most basic metrics to compare the automatic and ground truth labels of a test dataset. Finally,

this session presents the grading criteria for the students activity, so they can be aware of the problem requirements and their weights.

<STAGE 2> Selection of the concept: The students will decide what concept they want to detect from the image. The question will be posed in the introductory session and, after leaving some days for the students to discuss and explore the multiple optins, a solution will be taken. For example, the discussion can be open in an online forum during these days where everybody can make their proposal and express their reasons for it. The final decision is taken in class by voting. It is suggested to drive a two rounds voting, a first one with all proposed concepts, and a second one with the two most voted ones during the first round.

<STAGE 3> Data collection: The first step in any machine learning problem is collection and labelling data. Each team must prepare a collection of images generated by themselves and agreed to be published under a free license (eg. Creative Commons). Together with the images, teams must also provide an annotation of whether the images contain or not the concept of interest agreed in Stage 2. The dataset must be splitted in two parts, one for training and another one for testing. The train part is published by the teacher on an online repository so that the other teams can use it during the development of their system. The test part is kept secret by the teacher until the end of the activity, as it will be the data used for the final evaluation of the system. By collecting data from different teams, the final dataset is rich in variability, a desirable property when dealing with machine learning problems. The teacher must clearly indicate the data formats for both images and annotation files to avoid compatibility problems when combining data

from different teams. It is advisable to provide an example in order to avoid common formatting problems in the naming of the files and their contents.

<STAGE 4> Mock evaluation: A first round of evaluation results is programmed around halfway the total duration of the activity. This session aims at ensuring that the students are regularly working on the assignment and solving practical issues related to the live processing of data. At the beginning of this sessions, the students are provided with a subset of the training data and have a limited amount of time to run in class their classifiers and generate an automatic annotation. Then, they must use the automatic annotation and the provided ground truth annotation to generate the quality measures of their classifier. At this stage, their solution does not need to learn, so the classification parameters do not need to be tuned according to a training dataset. This simplifies the problem, and forces the students to focus on the software architecture and proper evaluation of the results. They are also asked to present on slides the work completed so far, as well as the results obtained in the first rehearsal. These slides should present the programming languages and environment chosen, the system architecture as a block diagram, the numerical results of the evaluation and the images that were misclassified. The teams should not provide details about the techniques they have chosen to solve the problem, this way the competitive nature of the activity is kept. At the end of the session, the professor can present more sophisticated measurements that are to be provided in the final solution.

<STAGE 5> Final evaluation: The final activity follows the same scheme as the mock up, but in this case the used data comes from the test datasets that were not available for

students until that moment. At this point the software must use the training data to auto-configure, and evaluate its performance by labelling the provided test set. The oral presentation must describe the image processing and classification techniques that have been adopted. These slides must be finished in class, as they must report about the results obtained on the new test dataset. In this oral presentation, the chosen techniques must be revealed in detailed and, if possible, proven with a live demo on a shared screen. The evaluation results must be provided in the previously harmonized metrics so that a performances of the teams can be objectively compared. The students are required to show the classification errors of their system and provide an explanation about these failures. Finally, the assessment of the same evaluation metrics on the same test dataset allows comparing whose solutions performed best. After the session, all team must publish their software together with a handbook. These publications are available to the rest of the teams, who can validate the authenticity of the announced results.

Table 1 shows a suggested scale for grading on a maximum of 100%, where the 50% mark corresponded to the minimal score to consider the activity as *passed*. The grades are communicated to the students split in four differents tasks, also explaining the interpretation given to each grade.

Table 1: Criteria for the grading of students

Weight	Description
60%	Mock evaluation session (30%) + Final evaluation session (30%)

	• 0% = the dataset was not processed 30 minutes after their publication • 10% = the dataset was processed, but no evaluation figures were provided and/or the oral presentation was confusing • 20% = the dataset was processed, evaluated and the oral presentation was correct • 30% = extraordinary performance, exceeding the basic requirements
20%	<u>Evaluation of the image classifier</u> • 0%=the minimum set of metrics were not computed. • 10%=all requested metrics were provided • 20%=additional visualization and/or evaluation strategies were obtained
20%	<u>Software and documentation delivered</u> • 0%=no working software or documentation were delivered • 10%=the results cannot be reproduced with the provided software and documentation • 20%=results can be reproduced with the provided software and documentation.

c) Results and/or conclusions

This activity was tested on a group of 23 students in the 3rd course of a Degree in Audiovisual Systems Engineering in the Engineering School of Terrassa (UPC). They self organized in 6 teams of 3-4 people. The diffreent stages of the activity were completed in 5 sessions spread in 6 weeks, as described and commented in Table 2.

Table 2: Proposed schedule and observations

Week	Session	Observations
1	Introduction and evaluation metrics	The presentation of the activity was a good way to link the theoretical contents that were being worked with slides, with a practical problem that the students were faced with. It was highlighted that the classification results were not as much important as a good methodology for the evaluation.
2	Selection of the concept	After an initial choice of "images containing bar codes", a professor from another course noted that this topic might be covered in a future lab. For this reason, the topic was changed to "day / night" images.
3	Data collection	Some groups generated new images for this task, some others just collected photos they already possesses, especially from holidays. There were some misunderstandings with the naming of the files, which forced a review of the activity instructions with more details and some examples.
4	Mock evaluation	Despite not being the final evaluation, the teams took this session very seriously and attended with great tension and concentration, much higher than in any regular lecture so far (see Figure 1). Some

Week	Session	Observations
		misunderstandings on the requirements were detected and solved during this trial.
5	No session	The group with best performance on the mock evaluation reached the professor to request help for a more complex system of evaluation, that had been introduced as optional. Other professors from the same department reported that they were consulted regarding the image processing skills necessary for the task.
6	Final evaluation	The level of attention was similar to the mock up session. All teams completed the basic evaluation protocol defined for the activity. In general, classification rates were high despite using an new and unseen test set. The oral presentations of the results were followed with interest as well as the demos (see Figure 2). Minor delays occured due to the execution of the demos or set up of the computer on the beamer.

Figure 1: Teams testing algorithms during the mock up session

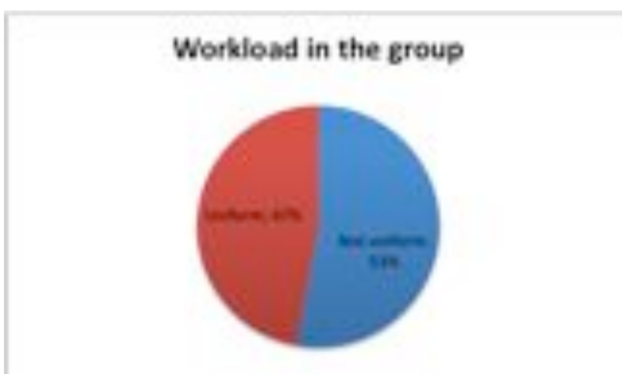
Figure 2: Live demo during oral the final evaluation session

The students were individually questioned during one semester after the course for a better understanding about the development of the activity. A total of 17 students were questioned, which represent the 74% of the whole class. They were asked about the workload of the activity and its distribution, both in sub-tasks and among the different members of the team. An overall opinion on the activity was also requested in quantitative and qualitive way. The results of this study are presented and analyzed in the following paragraphs.

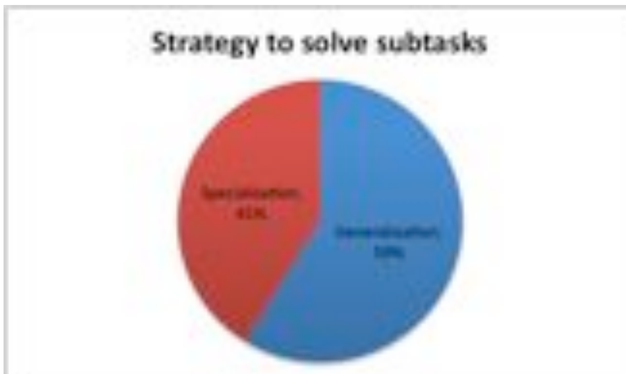
c.1) Workload distribution among team members

In this initial edition of the activity, no guidelines were provided regarding the internal organization of the groups. The first set of questions aimed at discerning how the roles and workload were distributed among the members of the teams.

Question 1: "How was the workload distributed among the members of your group ?"



Question 2: "What strategy did your group adopt to solve the different subtasks of the assignement ?"



The collected data can be completed with an additional question posed at the end of the course, where students were asked to privately evaluate their group mates with a simple *pass / not pass* qualification. In 5 of the 6 groups, all students gave a *pass* grade to their mates, but in one of the teams one student was graded as *not pass* by his/her colleagues.

The results indicate a diversity of cases among groups, with no special correlation between the uniform/non uniform and generalization/specialization approaches.

The *non uniform* cases, more than the half, are the most problematic result because the evaluation grade that students receive was uniform, equal for all members in the group. This means that some students were unfairly qualified, whether because they exceeded the expected workload to compensate the low dedication of others, or because their grade was over-estimated according to their actual contribution. It is difficult for the teacher to estimate the partial contribution of each group member without a huge increase of the evaluation effort. A possible solution for the situation could be based on a more defined co-evaluation among equals. The basic *pass / no pass* grades collected in the first edition, could be replaced by an estiation of every student about the workload assumed by each member of the group. This proposal could be privately delivered by each member, although they would be encouraged to openly discuss it and ideally agree on a common proposal for each group. This way, the teacher would obtain valuable data that would assist in a fair qualification of each individual.

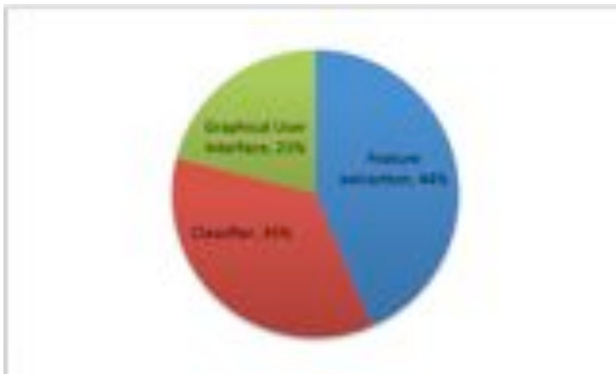
Regarding the adopted strategies for solving the subtasks, most groups chose a *generalization* approach where every student was involved in the different parts of the system. However, an important 41% of the students specialized in a sub-part. There was no indication from the teacher on the best strategy to solve the problem, although it is desirable that all students know about the whole system. The study plan includes *team working* as one desirable competence for graduates, which can be also developed in specialized distributions. Team working requires a co-operative approach and co-responsability on the results, but can also be fully achieved in a specialized framework. However, it is desirable that all students know about the whole system; not necessarily participating in the development of all parts, but at least being able to describe the design principles that govern them. This expected quality of the activity might not have been achieved during the first edition, but could be guaranteed by a simple modification of the mechanism for oral presentations. Instead of letting teams choose the person that would speak in front of the class to describe the work and results, future editions of the activity will require that any of the group members should be able to present the results. This way, every group will have an incentive to ensure that all of its components understands the complete system, with no need to really develope it. In fact, in most engineering companies the person responsible for the communication is different from the person who develops the product. In this activity, though, it is desirable that all participants develop and can present the final work.

c.2) Workload distribution among sub-tasks

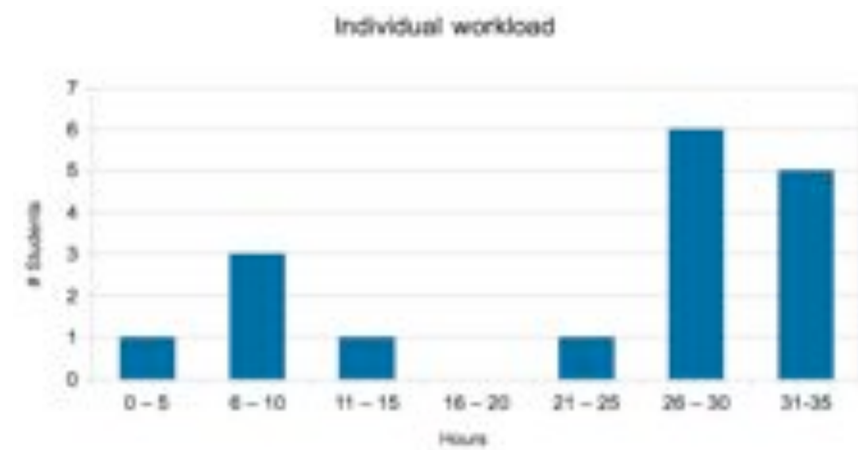
The classification of images involves the construction of a full system with different sub-parts. From the study plan perspective, the most important part corresponds to understanding the role of each sub-parts and being able to evaluate the whole system. Given the complexity of the task, it was important to guide students in focusing on those contents that corresponded to the course, and try to minimize their dedication on those parts that, despite necessary, are already addressed in other courses. For this reason, during the presentation session it was highlighted that the goal of the activity was to correctly evaluate a classification system, not building the best classification system.

Moreover, given that this was the first edition of the activity, it was also important to assess the workload in terms of hours of student work. In the current framework of a common European Higher Education Area (EHEA), every activity must be precisly quantified in order to adjust the contents of the study plans to the expected effort from the student side.

Question 1: "Provide an estimation of the dedication ratio to the main sub-parts of your project".



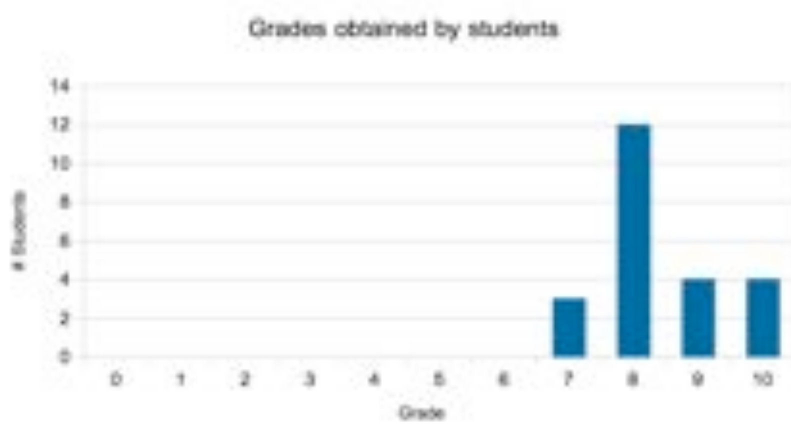
Question 2: "How many hours did you dedicate to this activity ?"



The definition of the sub-tasks introduced an unexpected result, which is the inclusion of the non-expected *Graphical User Interface*. In any point of the activity description it was requested to students to develop a graphical layer for their project, the only graphical material they needed to generate were the slides for the oral presentations. This part was probably generated by the students to help themselves to understand the behavior of the software, but it certainly too an important amount of workload, a 21%. In future editions of the activity, this component will clearly be described as not evaluated.

The two other main sub-tasks, *Feature Extraction* and *Classifier*, did match the expected outcome and were in fact described during the contents of the course. However there was a missing sub-task, which would have been the *Evaluation* part. The instructions of the activity invited students to completely separate the Classifier system from the Evaluation one. This split was not followed in any of the proposed designs, as they mixed the two systems in a single piece of software. Future editions of the activity will avoid this problem by introducing the cross-evaluation of systems, that is: instead of asking each group to evaluate their own system, the classification results will be saved in a file that will be delivered to the rest of the teams to be evaluated. This way, both *Classifier* and *Evaluation* parts will need to be developed separately, and it will be much easier to detect inconsistencies in the way that groups evaluate the systems. This separation will also facilitate the development of software, as these will be completely independent engines that can be completely developed in parallel.

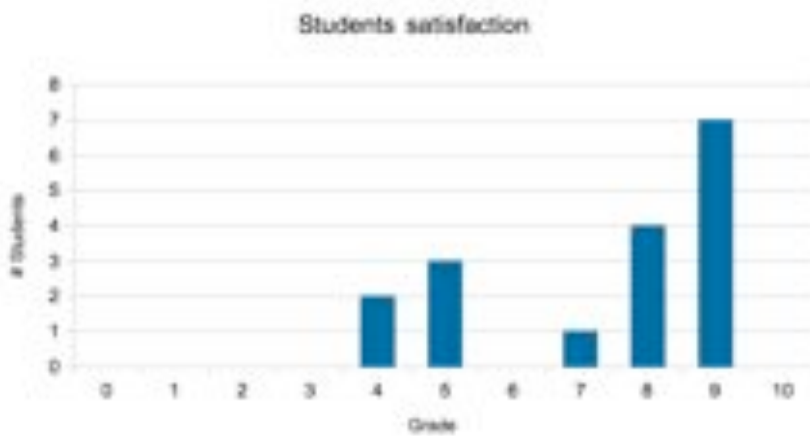
In terms of workload planning, the activity was expected to be completed in 25 hours of autonomous work from the student side. This estimation adjusted to most student dedication, as the average of dedication was of 25 hours per student. Results indicate that 5 students did not fulfill the expected workload, a situation that remained hidden during the course given the group-based nature of the activity. As discussed in the previous section, co-evaluation and preparing the oral slides should correct this bias. On the other end, 11 students worked more than the expected 25 hours in the activity, but never exceeding the 35. This additional dedication was translated in the high quality results obtained, that largely exceeded the initial expectations. The good results were translated into high grades for students. As previously commented, the group-based qualification was especially benefitial for the minority of students who did not accomplish the expected workload but who favoured from a high mark thanks to their team mates efforts. The following histogram shows the distribution of grades among students, with an average mark of 8,4 out of 10.



c.3) Overall evaluation of the actity

The problem-based learning approach adopted for this activity was aimed to boost the motivation of the students in learning the machine learning principles when applied to multimedia analysis. The third part of the questionaire tried to estimate the degree of satisfaction of students with the activity, as well as collecting their feedback to detect the strong and weak points of the activity from their perspective.

Question 1: "Grade your overall satisfaction on the activity with a scale from 0 (minimum) to 10 (maximum).



Question 2: "What is our opinion on this activity ?"

- I missed more guidelines about how to extract visual features.
- Amusing activity that proved useful to learn the contents of the course and discover new things.
- Too much workload for too little impact in the final grade.
- I liked this activity, I learned and it was amusing.
- It should have a higher impact on the final grade.
- More weight in the final grade.

The average satisfaction score was 7,32 and there were as many positive as negative comments. The positive comments match the expected results of the activity, which were exploiting the problem-based approach to teach students the parts of a classifier and how to evaluate it. Regarding the negative comments referred to the low relevance of the activity in the final mark, they referred to the 10% weight on the total course grade. In terms of workload, the 25 hours planned represented a 16% of the total workload of the course. There are a few correction factors that should be considered when interpreting these data. The first one, is that students unexpectedly decided to develop a graphical interface when that was not requested, and this part took them around 20% of their time. Without this additional work, their average dedication would have been around 20 hours, which would have corresponded to 13% of the total course. In addition, the topics covered during this activity were also requested during additional lab and problems exams, so a satisfactory job on this assignment would also have a positive impact on other scored activities. As a result, the assigned weight of the activity in the final grade is close to the amount of working hours planned in the course guide. On the other hand, it has been observed that in many cases students ignore the workload estimations when they plan their semester. This way, they tend to underestimate the expected effort to pass the courses and systematically blame the teacher for exceeding the amount of hours that they consider that a course should require.

To sum up, despite being the first edition, the results of the activity largely exceeded the expectations as most teams completed the designed goals and added some extra bonus, such as graphical user interfaces or more complex evaluation schemes. The practical and competitive nature of the activity were probably their best motivational arguments. As a negative result, the team-based approach may drive into specialization tasks that might prevent that all students complete all learning tasks. In the extreme, the team approach can also create unfair situations, where the workload in unevenly distributed among the components.

The evaluation of the activity is positive from the authors perspective, as it succeeded in introducing valuable new contents to the study plans in a fashion identical to the one adopted in academia and industry. Students learned how the skills on electronics and signal processing can be exploited with this third piece to build practical applications that can solve real problems. It is the will of the authors to increase the role of machine learning techniques in the study plans given their tremendous potential in the multimedia field.

6. REFERENCES

[Bishop06] C.M. Bishop, "Pattern Recognition and Machine Learning". Springer, Cambridge (UK) 2006.

[Doud99] D. Boud and G. Feletti, "The Challenge of Problem-Based Learning". Kogan Page Limited, London 1999.

[Nowak11] S. Nowak, K. Nagel and J. Liebetrau, "The CLEF 2011 Photo Annotation and Concept-based retrieval Tasks". Working Notes on CLEF 2011.

[Smeaton09] A.F. Smeaton, P. Over and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: A 5-Year Retrospective of Achievements". In Multimedia Content Analysis, Signals and Communication Technology, p. 1-24, Springer 2009.